The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration.

FINAL REPORT ON ONR GRANT NO. <u>N00014-02-1-0754</u>

PERIOD OF PERFORMANCE: <u>June 3, 2002 through December 31, 2003</u>

ABRIDGED VERSION FOR PUBLIC RELEASE

April 21, 2004

Principal Investigator:

Susan Opava, Ph.D.
Dean of Research and Graduate Programs
California Polytechnic State University
San Luis Obispo, California

Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
26 MAR 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration (Abridged Version)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
California Polytechnic State University Foundation Sponsored Programs Department Foundation Administration Bldg. 15 San Luis Obispo, California	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**364**	

TABLE OF CONTENTS

FINAL REPORT ON ONR GRANT NO. N00014-02-1-0754
PERIOD OF PERFORMANCE: 6/3/02 to 12/31/03

I.   Title of Project

The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration.

II.   Summary of Project

The mission of the California Central Coast Research Partnership (C^3RP) is to facilitate the exchange of knowledge and skills between the higher education sector and the private sector in San Luis Obispo County, and to encourage the growth of high-tech companies in the region, thereby enhancing economic development and quality of life. The partnership is a long-term plan to create a dynamic and self-supporting university-industry-government partnership that capitalizes on the strengths and mutual interests of the educational and technology-based business sectors. The plan recognizes the key role of higher education in preparing a highly skilled work force and transferring new knowledge to practical uses. The outcomes of this partnership, when fully realized, will be the expansion of current and the creation of new University technology R&D activities; the development of existing technology-based businesses and the creation of new ones; and the generation of opportunities for job training and research and development activities for University and Community College students and faculty in areas of interest to the Department of Defense and national security.

The project will eventually lead to the construction (with private financing) of a technology park on the California Polytechnic State University campus that will provide state-of-the-art space for private technology companies engaged in research and development activities, as well as a business incubator that will provide all of the support services needed by start-up, technology-based companies. At this stage, the project involves planning, analysis, relationship-building and pilot research projects related to development of the long-term partnership and its research foci.

III.   Relevance to ONR Objectives

A.   Relevant partners.

C^3RP represents a coalition of educational institutions, local, state and federal government, and private businesses that have worked together in unprecedented fashion to advance the common goals inherent in the proposed university-industry partnership. The current partners in the project and their contributions include:

California Polytechnic State University
- o   committed the land for the project, valued at ~$1.5 million
- o   provided assistance in financial management of the project

- o contributed $90,000 for a pre-feasibility study by Bechtel Corporation
- o committed several hundred thousand dollars of in-kind contributions of senior management time and effort over several years; continues to do so
- o invested ~$800,000 in efforts to raise additional funds for the project

GEO, International (GEOgraphic Network Affiliates, International; a private company)
- o works *pro bono* with C^3RP on communications aspects of the project

CENIC (Corporation for Educational Network Initiatives in California; association of Internet2 universities in CA)
- o works with Cal Poly and GEO to further the goals of the IEEAF (see below), goals that will directly benefit C^3RP
- o CENIC and AARNET, Australia's Academic and Research Network (http://aarnet.edu.au) are working on a collaborative effort with the fiber-optic carrier, Southern Cross, to implement a trans-Pacific high-speed network connection (10 Gigabits/sec) that will connect through the Cal Poly campus.

City of San Luis Obispo
- o in partnership with Cal Poly has developed a carrier-neutral fiberoptic ring around the city (see below for detail)

IEEAF (International Educational Equal Access Foundation; established by GEO and CENIC)
- o is currently securing donations of virtual and physical communications assets in at least 37 countries; some of these fiber-optic assets will directly benefit this ONR project

National Science Foundation
- o is working with GEO, IEEAF, CENIC and other universities in the United States to promote the goals of these organizations to develop low-cost fiber-optic networks for the benefit of educational institutions, non-profit organizations and local communities.

Efforts are ongoing to secure new partners, including:

- National Center for Research Resources of the National Institutes of Health
- National Guard Bureau (federal)
- California National Guard
- National Interagency Civil-Military Institute (federal)
- Governor's Office of Emergency Services of the State of California
- Major corporations

B.   Strategic location.

New technological developments in San Luis Obispo related to the intersection of major undersea and terrestrial fiber-optic cable networks on the central coast of California have provided an exceptional opportunity to focus the research partnership initially on the emerging technology area of global telecommunications. Linking the research partnership to this burgeoning field has enabled us to recruit many of the important partners listed above and will provide the strong affiliation between the educational and private sectors that is the foundation for success in university-related technology parks. It will also open new opportunities for federally supported research and development projects.

C.   Relevant R&D focus.

Global telecommunications, particularly wireless, "over the horizon" communications, is fundamental to many current and developing defense strategies, including the ability to respond effectively to various forms of terrorist activities (including biological and chemical warfare), and in situations that require large-scale humanitarian assistance and disaster relief – areas which have become far more critical in the wake of the September 11, 2001 terrorist attack on American soil. It is also fundamental to the development of the decision-support systems that underlie and are key to these strategies. Cal Poly's Collaborative Agent Design (CAD) Research Center is the architect and lead developer of one of the first such systems: IMMACCS (Integrated Marine Multi-Agent Command and Control System), with JPL, SPAWAR Systems Center and the Stennis Space Center (NRL) as the other principal team members. Because of this, Cal Poly and the CAD Research Center are poised to take the lead in the continued development of complex decision-support systems and to develop a center of excellence in this area that will place Cal Poly at the forefront of the field. The C^3RP project has provided and will continue to provide support for the CAD Research Center to continue, and expand, its work on on-going projects like those completed for the MCWL, such as IMMACS, and the joint MCWL-ONR ELB (Expanded Littoral Battlefield), ACTD (Advanced Concept Technical Demonstration) project.

The scope of the C^3RP project, however, is much broader than simply the continuation of current projects of the kind described above. Command and control technology is developing rapidly and will have nearly universal applications. The basis of this technology is the ability to store "information" rather than "data", a concept that is fundamental to the capacity to utilize that information to support complex decision-making. IMMACS is one of the first of such systems to use this concept, and has led the field, but the bringing together of commercial and University research and development through C^3RP will lead to an explosion of applications and technological advances in this arena.

In addition to telecommunications, many other research areas have been targeted for development through C^3RP. These have been selected because Cal Poly has demonstrated or developing strengths in these areas. They include:

Agricultural and environmental biotechnology	Telecommunications technologies
Aerospace technologies	Photonics
Rendering, animation and modeling	Graphic communication technologies
Software engineering	Nano- and micro-technology
GIS and GPS applications in agriculture, biology, architecture and engineering	Computer engineering (small systems, peripherals, custom applications)
Network hardware and software technology	Surveillance, logistics, command & control support systems
Data processing	
Remote sensing	e-Learning software development
Polymers and coatings	Bio-engineering
Electroptics	Robotics
Power generation and distribution	Magnetic levitation (transportation)
Biological risk assessment and detection	Risk assessment and prediction of fire terrorism
Protection of the nation's food and water supplies	Environmental hazard assessment and detection
Transportation disaster management and control	Seismic research

The majority of these research areas mesh with strategic initiatives identified by the Office of Naval Research. Research and advanced development projects in these strategic areas are being developed, often in collaboration with industry partners identified through C^3RP and recruited to join the partnership.

D.   University-industry-government partnership.

The primary focus of this initiative is to forge a strong link between private sector R&D and University applied research to speed the development of new knowledge and the transfer of technology to the public and private sectors. San Luis Obispo has become a draw for technology businesses (with a heavy concentration of software development companies) from both the LA Basin and Silicon Valley. For example, SRI (Stanford Research Institute), International operates a "software center of excellence" in the city. Branches of major corporations are also located nearby, for example, Sun Microsystems, AutoDesk, and Sunbay Software. Lockheed-Martin has a research and development group in nearby Santa Maria. At least one small local business is currently working on defense contracts. Two local companies manufactured critical components for the Mars rovers. Also located on the Central Coast are several government entities with whom we are working to advance the goals of the C^3RP project. These agencies have been listed on p.2 and include NICI (the National Inter-agency Civil-Military Institute), as well as state entities whose mission includes anti-terrorism training/response and response to large-scale emergency situations (California National Guard and the Governor's Office of Emergency Services).

E.   University strengths.

Cal Poly is a State university that has achieved national distinction as a polytechnic university, with engineering and computer science programs ranked among the very best undergraduate programs in the country. It is moving with vigor into the 21st century and these characteristics have led it to orchestrate the research partnership effort and the consortium of partners proposed herein. Cal Poly also has affiliations with CSA (California Space Authority) and with Vandenberg Air Force Base, where it has offered an M.S. in Aerospace Engineering by distance learning. The high bandwidth that will be

associated with the eventual physical site selected for the partnership will allow Cal Poly to offer many more academic programs by distance learning to remote locations. In particular, we will have collaborative agreements at cable-head locations around the world (including Asia and Europe) that could make our programs available to military personnel stationed almost anywhere in the world. This could be tied into training programs for ONR and, if desirable, to training directly related to research projects.

In summary, the California Central Coast Research Partnership offers an unprecedented opportunity to take advantage of a confluence of factors, including existing and potential relationships, fortuitous and unique technological and economic developments in the region, the particular strengths and expertise of the CAD Research Center, and a meshing of the research and development interests of the University, the Office of Naval Research, and the private sector. C^3RP is the vehicle for fully realizing the benefits of these common goals and synergies.

IV.   Summary of Results During the Period of Performance

A.   General.

This project was originally funded through an award from ONR in FY '02, covering the period September '01 through December '02. Accomplishments during this first award period were described in a detailed report to ONR in March '03. The current award began June 6, 2002 and ended December 31, 2003. Progress on this award is summarized below. Detailed reports are presented later in this document.

An overview of accomplishments to date on the project follows:

- Research carried out by the **CADRC (Cooperative Agent Design Research Center)**, of particular interest to ONR and the Marine Corps, was again funded. A detailed report on this project, Tools for Creating Mappings between Related Ontologies, is provided later on in this document.

- Research carried out by the **CADRC (Cooperative Agent Design Research Center)**, of particular interest to ONR and the Marine Corps, was again funded. A detailed report on this project, Tools for Creating Mappings between Related Ontologies, is provided later on in this document.

- The project director and manager and numerous faculty met with Mr. Fred Belen, in his capacity as a consultant for ONR. Mr. Belen made two visits to Cal Poly (8/19 and 9/30-10/1/03) to discuss progress. Mr. Cliff Anderson, director of basic research for the Marine Corps, accompanied him on the second visit. During that two-day meeting, they met with dozens of faculty who discussed their currently funded projects as well as proposed projects in areas of interest to the Marine Corps and ONR.

- **New research** has been developed and some research has been continued, including the following projects, many with industry collaboration. They include topics highly relevant to defense and national security. Detailed reports of the results of these projects are presented in Section IV. B. of this report.

  - Design Methodologies for Analog/Mixed Signal VLSI Systems Applied to Infrared Focal Plane Arrays
  - Optical Detection of Airborne Micrometer-sized Particulate Matter
  - Non-Linear Aircraft Simulation for Aircraft Test and Design
  - Heat Collection Element Glass to Metal Seal Redesign for Solar Power Generation Plant
  - Range Sensing and Real-Time Registration
  - Development of an Autonomous Tactical Reconnaissance Platform
  - Fast, Remote Infrared Spectroscopy of a Rocket Plume
  - Correlation of Milk Composition and Fouling with Biofilm Formation and Microbial Spore Production in Heat Exchangers
  - Development of a Rapid Method for Detection of Early Spore Generation Using Molecular Probes and Electro-chemical Luminescence
  - Access to the Central Coast's Marine Environment through a Real-time/Archived Data Interface
  - Network Performance Research
  - A Viable Remote-Sensing Multispectral/Hyperspectral Compression with Region of Interest Prioritization Capability
  - Construction of Tetrahedral Photonic Bandgap Crystal: Demonstrating Three-Dimensional Self-Assembly Using DNA Linkage
  - Development of Launch Interfaces for the CubeSat Program

- Four projects received small amounts of funding to explore feasibility or to complete work on projects previously funded:

  - Development of Field Rechargeable Gas Mask Filters
  - A Homogenization Approach to the Analysis and Design of Composite Systems
  - Water Treatment with Manganese Oxide Biofilms
  - A Fault-tolerant Flight Controller Design

- Significant enhancements have been made in **information technology infrastructure.**
  - **Internet2** connectivity was applied for, approved, and acquired for the campus in November 2001, to support current and future research efforts. Internet2 membership was continued and fiber-optic connectivity was extended this year between Cal Poly and Cuesta College, enabling the community college to have access to Internet2 for the purpose of collaboration with Cal Poly.
  - To ensure our flexibility to access fiber-optic carrier points of presence and also to promote connectivity to City resources (e.g. City/County

Library), Cal Poly entered into a **partnership with the City of San Luis Obispo** to install fiber-optic cable within City conduit. The work has been completed and the enhanced access it ensures will be critical for certain research applications, as well as for the viability of the technology park. **This project was accomplished without use of ONR funds.**

- o CENIC (see above) and AARNET, Australia's Academic and Research Network (http://aarnet.edu.au) are working on a collaborative effort with the fiber-optic carrier, Southern Cross, to implement a **trans-Pacific high-speed network connection (10 Gigabits/sec)**. The California-based landing site will be in San Luis Obispo County and will connect into CENIC's high-speed network at its network point of presence located on the Cal Poly campus. Cal Poly's geographic location, its ability to act as a "hardened" facility to support critical network backbone equipment, as well as the existing fiber-optic infrastructure built out in partnership with the City of San Luis Obispo, has positioned Cal Poly to be the access point for the high-speed network corridor to Australia's research environments and potentially other Eastern Pacific Rim research networks. **This project was accomplished without use of ONR funds.**

- **Several pilot projects** were funded to explore **new uses of Internet2** for research and education. Three presentations on the results of these projects were competitively selected and presented at the program of the annual Internet2 conference in October 2003. The projects are:

  - o VLSI Design and Test with Internet-Based Software and Hardware Tools
  - o Internet Map Server for Geographic Information: Further Development
  - o I2-Based Remote Viewing System
  - o Collaborative Research in Next-Generation Embedded Microprocessor Designs
  - o Video-capture, editing and production on I2
  - o End-to-End Performance Testing

- A **database of technology–based companies** that are potential partners in the project and research collaborators was developed and is continuously updated and expanded.

- **Relationships** have been established with new technology companies that are potential research collaborators, including: Interstate Electronics, Reson, DeRoyal, AgSCAN, Strategy Insight, and Dynamic Technology.

- The web site for the project (www.c3rp.org) was completely redesigned. The updated site presents C^{3}RP as an interface between Cal Poly and business/industry for the purpose of facilitating R&D relationships. It presents Cal Poly's special strengths in applied research and encourages collaboration of various kinds. The technology park is described as a future piece of our efforts to partner with industry in research and development. A link entitled "For Business

and Industry" will take visitors from the new "Research" page on the Cal Poly web site (www.calpoly.edu) to the C^{3}RP web site.

- An **industry patent-donation program**, established with SAIC (Scientific Applications International Corporation) resulted in new patent donations. These patented technologies can be used to develop new lines of University research.

- A sub-group of the University President's Cabinet (executives and former executives from major US corporations) has been formed to provide on-going advice and guidance for the project. The C^{3}RP project was presented to the entire President's Cabinet in April '03. Subsequently the project leaders met with the sub-group of the Cabinet several times and continue to seek advice from this group on an on-going basis.

- Efforts continue to develop industry partners in the **biotechnology sector** for the purpose of developing research and training activities in this field. To this end we have been working with the Central Coast Biotechnology Center in Ventura, CA, with two local community colleges, and with several biotech companies, including Amgen, Baxter, FzioMed, Genentech, Promega BioSciences, Hardy Diagnostics, and Santa Cruz Biotechnology.

- The project's leaders have continued to work with other private and government (including HUD and EDA) partners to advance the project and to attract research collaborators and support.

B. Detailed Research Reports

Following are detailed reports of the technical results of research projects carried out this year. They are presented in the following order:

1. Results of CADRC research
2. Results of fully-supported projects
3. Results of exploratory or follow-up projects
4. Results of Internet2 pilot projects

Tools for Creating Mappings Between Related Ontologies

Jens G. Pohl, Ph.D.

Collaborative Agent Design Research Center (CADRC)
Cal Poly State University

Topic: Tools for Creating Mappings Between Related Ontologies

Abstract

In previous work in the area of information-centric systems for disaster management planning, investigators in the Collaborative Agent Design Research Center (CADRC) identified the likelihood that such planning would include the use of knowledge resources from multiple organizations. In many cases, it is unreasonable to assume that all organizations would adopt identical representations of information. The likelihood of multiple representations leads to a requirement for translation, or *mapping* from one representation to another, using methods that retain the context and relationships between objects as they are communicated among diverse systems. Our investigation examined current methodologies for constructing such mappings using various approaches, and implemented two example mapping techniques.

Introduction

In disaster management, as in many other types of problems that fall into the general category of resource management, planning must be based on information from various automated and semi-automated systems, including systems from multiple organizations. Such systems are generally built in isolation from one other, and seldom attempt to take advantage of data or information representations that may exist in systems that deal with related tasks, even when these systems are used within the same organization. This heterogeneity of representation is one of the key factors preventing information sharing at the system level.

Past efforts to implement communication among systems have often focused on the data level of representation. This level is only concerned with defining names and types of individual fields in the data, along with simple relationships such as tuples and foreign keys (as in a relational database). Data transmitted from one system to another under these circumstances can be correctly received, but the more complex relationships among these data elements have been lost. Accordingly, the receiving system will be unable to treat the input as information, except in some predetermined, hard-coded manner. As a result, as the number of systems attempting to communicate grows, each participant becomes increasingly complex and difficult to maintain and extend. Heterogeneous distributed systems based on point-to-point data models tend to be brittle and unresponsive to change.

Another approach has been to define a single data or information model for the entire system of systems. In the single model approach, each participating system is expected to provide, at a minimum, an application interface that can send and receive data or information based on the common model. In some cases, each system may be expected to implement the common model internally. The rationale behind such a requirement is that the storage used by each system (e.g., a relational database management system) should be directly accessible by other systems, without using the originating system as an intermediary.

2

The success or failure of any implementation of the single model concept largely depends on the ability of a single organizational entity to enforce the use of the model across all the participating systems. There are circumstances in which this is possible: for instance, organizations that are dependent on the entity that is attempting to require a single model are likely to implement that model rather than lose their dependency. In many cases, however, the single model idea has proven problematic in implementation even within individual organizations. Cross-organizational attempts tend to flounder, often failing even to reach agreement on the model, much less to accomplish implementation.

In a situation where each system, or at least each organization, is likely to have its own data or information model, how can interoperability be achieved? For systems that are exclusively data-centric, this is a problem with no satisfactory answer. Such systems must rely on exposing a data model, with the attendant difficulties of brittleness and resistance to change discussed above. For information-centric systems (e.g., ontology-based applications), on the other hand, there are some approaches that offer opportunities for success in interoperating in a heterogeneous environment.

Regardless of their internal representation, each system can expose its information model as an ontology, using a standard language such as DAML+OIL [daml] or OWL [McGuinness03]. Such ontologies include relationships among concepts in the model, and allow automatic reasoning about the model itself. When a system provides its model in this form, other systems are enabled to view that system's information in the same context that the system itself does. This is a large step forward in the process of providing interoperability at the information level.

Exposing each system's own information model in a form that allows automatic reasoning is not, however, the complete solution to interoperability. There is still the sizable problem of translation from one information model to another. For this, systems need to determine not only the context of the system providing the information, but also the relationships between that model and their own. Given a description of the relationships that exist between models, it becomes possible for systems to automatically generate translators between multiple models. This is the beginning of truly flexible cross-system interoperability.

However, the issue of constructing the relationships between models still remains. These "ontology mappings," as they are referred to, can be complex to construct and frequently require intimate knowledge of the models used by heterogeneous systems. This requires a level of expertise that is not commonly available, particularly when the systems were built by different organizations. Creating ontological mappings is a specialized skill, and yet these mappings are vital to allow information-centric systems to communicate.

A number of projects are attempting to provide tool sets for creating mappings, and for using those mappings to automatically merge information from one ontology into another. The following section of this report examines several of the leading examples in these fields.

Following this examination of the current state of the 'mapping and merging' area, the report discusses the work performed by the Collaborative Agent Design Research Center (CADRC), Cal Poly, San Luis Obispo, to support one kind of mapping between ontology models, and a related project designed to allow users to develop and evolve ontologies

3

expressed in DAML+OIL. Finally, the report suggests possible future work and draws some conclusions. A list of references follows, and appendices listing the ontologies and axiomatic semantics of the ontology development project conclude the report.

Survey of the Field

Due to the level of importance of techniques for defining mappings between ontologies, as well as for automatically merging ontologies based on those mappings, this is a very active research area. What follows is necessarily a small, but hopefully representative, sampling of the work being done in the field.

Wache [WACHE01] lists three possible results of an ontology mapping system: a single global ontology; multiple ontologies with mappings from one ontology to another; and, hybrid methods, which involve multiple ontologies with a shared vocabulary.

The single ontology concept requires the semantic representation of a number of heterogeneous information sources to be mapped to elements of a global ontology including all concepts and relationships of all sources. This seems an attractive approach since, given such a mapping, integrating information from multiple sources becomes reasonably straightforward. However, a single ontology that faithfully represents the ontological commitments of all sources can be difficult to achieve. Additionally, this approach can have a high maintenance cost, since changes to the representation of any source may require disproportional changes to the global ontology.

In the multiple ontology approach, there is a single ontology for each information source. To overcome the use of multiple vocabularies, a representation of the relationships between terms in each vocabulary is defined. This idea seems attractive because it avoids both the ontological commitment and the maintenance problems of the single ontology approach. But multiple ontologies carry their own penalties: first, there is the well-known combinatorial explosion problem, as each individual ontology must map to each of the others; and, second, the mappings themselves require significant effort to produce, since each ontology may use a completely separate vocabulary from each of the others.

The hybrid approach seeks to mitigate the problems of the other two. Here, the information model of each source is mapped to a separate ontology, but the concept vocabulary of all models is drawn from a single global vocabulary. The terms of the global vocabulary serve as primitives for constructing the individual ontologies, through combinations based on a given set of operators. Because all the ontologies share this common base vocabulary, mappings between the ontologies are simpler to construct than in the pure multiple ontology model. On the other hand, it may be difficult to define an ontology representing an existing information model, in terms of the global vocabulary. And, if an information source has an explicitly defined ontology, it is unlikely that this ontology can be used directly in the hybrid model. In this case, an entirely separate new ontology, compatible with the global vocabulary, must be created for the information source.

From this discussion, the difficulties of creating ontological interoperability for heterogeneous systems can readily be seen. None of the three approaches provide a satisfactory solution in the general situation. This difficulty is implicit in the problem under study: there is no methodology for implementing a level of understanding between

4

systems with differing models of the universe of discourse that does not have severe drawbacks.

This does not mean, of course, that solutions cannot be found in specific cases. In many domains, the vocabulary used by subject matter experts is reasonably standardized. Information models that attempt to represent the knowledge of these experts, then, are likely to share vocabulary terms. This may make the mappings in the multiple ontology approach feasible, as well as assisting in creating both the shared vocabulary of the hybrid model and the hybrid ontologies themselves. In other words, the lack of a satisfactory general solution does not prevent moving ahead in individual cases.

Many of the projects in semantic mapping are concerned with the statistical analysis of representative (or *instance*) documents. Xiaomeng Su [SU03] reports on a technique based on information retrieval (IR) theory, in which concepts from one ontology can be seen as queries against the concepts in another ontology. The result of each "query" is the set of concepts that best relates to the initial concept. This technique does not completely automate the creation of ontology maps, but instead is intended to supply input to a human effort to determine the relationships between two ontologies.

In Su's technique, concepts in an ontology are enriched by adding a "feature vector" which is derived from documents that represent instances of the concept – that constitute, in other words, the *extension* of that concept. Once feature vectors have been calculated for each concept in a pair of ontologies, each vector from one ontology can be compared to each vector in the other. The two vectors that are most similar are likely to be closely related, if not identical.

From [Pekar02], we learn that a "feature vector" consists of the vector of "…words appearing within some delineation around the targets word," along with their counts. This is also known as the "kNN" approach, for "k Nearest Neighbors." A variation is the "Rocchio Algorithm" [Joachims97], which also takes into account the "inverse word frequency" across a collection of documents. The inverse word frequency will be low if the word appears in many documents, indicating that the value of the word as an index concept in a particular document is also low. This approach focuses, then, on the words that discriminate one document from the others in the collection.

The complete Rocchio algorithm then goes on to include "relevance feedback", which allows users to train the query engine to optimize the search. This also could be used in ontology mapping (creation) systems, to allow for human intervention in search results.

The application of this technique to ontology mapping involves calculating feature vectors for a set of documents, based on two different ontologies. The feature vectors can then be compared, looking for the vectors that are most similar to vectors in the other ontology. One frequently used means of comparison is to calculate the cosine correlation: the dot product of the two vectors, divided by the product of their lengths. This results in the angle between the two vectors in n-dimensional space, where 'n' is the number of terms in the vectors. The smaller the angle, the greater the similarity between terms. This process is based on the belief that the meaning of a term is defined by its context, and the context can be inferred from the other terms that are frequently used in conjunction with the term under consideration.

5

The user then must determine whether the similarity is real or merely a statistical anomaly. In the system described in [SU02], there are a number of mapping assertions that can be made concerning the relationships between terms in different ontologies. From these terms, a mapping system can automatically execute queries defined in one ontology, to retrieve documents that were originally described in terms of another.

These statistical analysis techniques are especially useful in two important situations. First, the similarity measure over an existing set of documents can be used as a basis for defining ontologies for those documents after the fact. When all the documents describe aspects of a single domain, this approach may not only reveal important concepts, but may also suggest relationships between them due to contextual similarity. This can be very important for categorizing a large set of documents semi-automatically.

Second, this type of ontology mapping can also be useful in providing interoperability between information portals, each of which may serve out documents based on a different categorization of terms. Once a mapping has been established, it becomes possible for a user to query using the local ontology, yet retrieve documents from other portals, without requiring the user to understand multiple information models.

Statistical analysis models have been primarily used, as indicated above, to categorize collections of documents. However, it may be possible to apply these techniques to the analysis of multiple explicitly defined ontologies by comparing fields of objects that are defined in accordance with one or another of the set of ontologies under investigation.

[Doan02] describes the GLUE system for finding mappings between ontologies. GLUE uses machine-learning techniques to find the most similar concepts in each ontology. GLUE uses multiple strategies, some dealing with instances of ontologies, and some dealing with the definitions of the ontologies, for learning about similarities.

The GLUE technique is built around the assumption that similarity between concepts can be calculated using the "joint probability distribution" of the concepts. By this is meant the probability that a given object is an instance of both class A and class B, the probability that an object is in class A, but not class B, and so on. One example of such a similarity is the Jaccard coefficient, which is the probability that an object is in the intersection of two classes, divided by the probability that an object is in the union of the two classes. This measure relates the number of instances that the two classes have in common, to the total number of objects in both classes. It is argued that the larger that ratio, the more similar the two classes are.

In order to calculate the Jaccard coefficient, GLUE must solve the following problem: given a number of objects, each of which is classified as either an instance of class A or an instance of class B, how can we determine which of those objects is also an instance of the other class? GLUE uses a variety of machine learning techniques to perform different types of analyses, and then joins the results to create a single classifier for each class. This classifier is then applied to the instances of other classes to determine which objects are also members of the class that the classifier represents.

There are two learners presently used in GLUE, although others would certainly be possible. The first is the content learner, which treats each object's attributes and their values as the object's "textual content." The probability that a given object is an instance

6

of a given class, given the instance's textual content is calculated using a "Naive Bayesian" approach. This approach treats the textual content of an instance as a bag of tokens. Each token is then used as input to a calculation of the relative frequency of appearance of this token as part of known instances of the class, compared to the total number of tokens appearing in all known instances.

The second learner used in GLUE is the name learner. The name learner uses the full name of each instance as input to a similar calculation as used by the content learner. The "full name" signifies the concatenation of all concept names in the taxonomy leading to this instance. From the results of the study undertaken to determine the accuracy of mapping, it seems that the **name learner** is much less significant than the **content learner**.

The results of the two learners are combined using the **meta-learner**. The meta-learner weights the output of each of the other learners to produce a combined probability that a given instance belongs to a given class. The weights are assigned manually, but, according to [Doan02], it should be possible to generate the weights through other machine learning techniques.

GLUE focuses on the taxonomy portion of an ontology (i.e., the hierarchical parent-child relationship) and uses several similarity measures to find the correct placement for a given class C in a taxonomy. These similarity measures include: (1) finding the most similar concept to C; (2) finding the "most specific parent" – the class that is the "most specific superset" of C; and, (3) finding the "most general subset" – the class that is the nearest child of C. The result of this analysis determines the most appropriate position for C in taxonomy T.

There appear to be certain limitations in the classification scheme used by GLUE. First, the algorithm works only on the taxonomy aspect of ontologies, but it seems that relationships other than hierarchical would also be important in determining mappings. Second, the process seems to assume that there will be a one-to-one mapping between concepts in one ontology and similar concepts in another, ignoring the possibility that partial matches (i.e., some attributes of one class matching some attributes of another) are also likely. However, the authors state that GLUE has been used successfully to map several real-world ontologies, so it may be that in practice these objections are not as serious as they might appear in theory.

The two approaches discussed above are primarily based on statistical analysis. Another class of systems for constructing mappings between ontologies focuses on interaction with the user to determine the relationships, with the system providing some level of support and guidance. The next two systems that will be discussed, MAFRA and Chimaera, are used to exemplify this approach.

While the previous systems focused entirely on automated discovery of semantic mappings, [Maedche02] discusses MAFRA, a framework that intends to cover all aspects of constructing mappings, including the execution phase in which instances of classes in one ontology are transformed into instances of classes in a second ontology. MAFRA also seeks to address the problem of the independent evolution of ontologies. Perhaps most importantly, MAFRA acknowledges that separate ontologies are the result of agreements among a community of human users, and that mapping between two such

7

ontologies is a cooperative effort among human beings. The results of any automated process of defining the relationships among concepts from diverse ontologies must be submitted to each community for ratification. Thus, MAFRA includes tools for collaborative consensus building.

The output of MAFRA is a set of "…semantic bridges", which contain "…all necessary information to transform instances of one source ontology entity to instances of one target ontology entity." The bridges themselves are defined by a meta-ontology known as the "Semantic Bridging Ontology," or SBO. The SBO includes such concepts as Service (an external resource that performs some aspect of transformation), Condition (which describes when to use a specific semantic bridge), and Composition (descriptions of combinations of semantic bridges), among others. This ontological approach to the definition of mapping is a significant contribution over and above the statistical methods described previously.

Another complete environment for creating ontology mappings, for merging ontologies and for general ontology design, is Chimaera [McGuinness00]. Chimaera is intended to solve problems related to merging very large knowledge bases, to infer new knowledge from the result, and to support semantically enabled searches across information from all sources. The initial knowledge bases came from the Defense Advanced Research Projects Agency's (DARPA) High Performance Knowledge Bases (HPKB) program [Pease00]. The knowledge bases in HPKB were prepared by various authors and were at diverse states of completeness. As a result, merging these ontologies involved identifying and adding missing relationships as well as standardizing terminology and usage.

Chimaera is designed to support ontology designers and to provide tools for testing and evaluating ontologies. [McGuinness00] describes the two major tasks of Chimaera as follows:

> "…(1) to coalesce two semantically identical terms from different ontologies so that they are referred to by the same name in the resulting ontology, and (2) to identify terms that should be related by subsumption, disjointness, or instance relationships and provide support for introducing those relationships."

Chimaera is primarily concerned with presenting ontologies to a designer in a way that enhances his or her ability to create a merged ontology. As such, Chimaera's reasoning about possible relationships across ontologies appears to be limited to considering similarities of the names of terms, and to a number of heuristics for finding candidate points for reorganization of the taxonomy. It is interesting, however, that even with these limited forms of assistance, Chimaera proved to be significantly more efficient than ontology editing systems that did not include tools for assisted merge.

Another important concern in ontology mappings is the requirement to use the mappings once they are created. Another system, OntoMerge, shows one method of driving translation through inferencing across ontologies using mapping rules.

OntoMerge [Dou03] supports merging for the purpose of translation. As the paper by the researchers puts it, their "…focus is on formal inference from facts expressed in one ontology to facts expressed in another." The OntoMerge process initially "merges" the two ontologies simply by taking the union of all terms and axioms in both ontologies.

8

The true definition of the merged ontology is formed by constructing a set of *bridging axioms* that expresses the relationships of concepts in one ontology to concepts in the other. The central purpose of OntoMerge is to allow queries, expressed in terms of one ontology, to be performed across knowledge bases that may be defined using other ontologies.

OntoMerge considers three possible types of semantic mismatches. First, the same term may appear in different ontologies, but the axioms associated with the term identify different sets of instances. Second, properties of a term may have the same name, but different meanings. Third, similar concepts in different ontologies may inherit different properties from their super-classes.

While OntoMerge assumes the existence of mapping rules, it does not appear to include tools either for automatically generating such rules, or for assisting human ontologists to create them. The rules themselves are expressed in a "strongly-typed first order logic language" that can be used directly by an inferencing engine to reason across sets of facts expressed in different ontologies, by translating queries and facts from one representation to the other as a result of the inferencing process.

It would appear that to be completely useful, OntoMerge would need to be paired with a toolset for extracting candidate mapping rules and for supporting human experts in creating the complete mapping. Without some form of automated and intelligent support for rule construction, it is unlikely that something like OntoMerge would be useful on a large scale. Given a rule set, however, OntoMerge is a good example of what is possible using inferencing to accomplish translation.

The statistical analysis-based methodologies for constructing ontological mappings rely heavily on the existence of a large corpus of documents or of large amounts of instance information relating to the ontologies under consideration. For our work, we were constrained by our intention to focus on the ontologies of existing or proposed projects in the CADRC and at CDM Technologies. Due to the fact that the instances produced by existing systems are not in a format that lends itself to this type of statistical analysis, and that proposed systems have no instance information, our project was limited to other types of processing. The use of statistical methods is likely to be a significant part of future work.

Discussion of work performed

XMI to DAML: Our initial exploration in creating mappings between ontologies focused on mapping from ontologies represented as object models in the Unified Modeling Language (UML) [Booch99] to an equivalent representation in the DARPA Agent Modeling Language (DAML) [Oullet02]. As a key part of the Semantic Web [Berners-Lee01], DAML and its successor, OWL, are expected to play an increasingly important role in expanding the influence and utility of ontology-based software systems. It must be expected that mapping to and from DAML or OWL will be an integral part of achieving interoperability among web-based systems.

UML has an XML representation named "XML Metadata Interchange" (XMI), and there is also an XML serialization format for DAML ontologies. This simplified the physical translation problem. By mapping the same ontology from one representation to another,

we were able to produce this first attempt at a more general level than would have been the case in mapping between two disparate ontologies.

For our test case, we used an ontology from one of CADRC's existing systems as a starting point. IMMACCS (Integrated Marine Multi-Agent Command and Control System) [Pohl 2003] is a complex distributed system connecting clients and servers from multiple vendors. For the purposes of this paper, the key point about IMMACCS is that the domain of knowledge is defined as on ontology in the Unified Modeling Language. As part of the system development process, the domain model is exported as XMI, and software generation tools create Java classes, CORBA IDL files, and a representation that can be used as input to an inference engine.

By creating a mapping between XMI and DAML, we hoped not only to gain experience in mapping ontologies, but also to show the feasibility of incorporating such automated mapping into an established development process. For this project, we used the Jena API for processing RDF and DAML statements [McBride03]. Jena allowed us to create DAML objects in memory and then to serialize them to the XML representation of DAML when processing was complete. We also used the Novosoft UML (http://nsuml.sourceforge.net/) API package for loading and manipulating the input XMI file. Thanks to Jena and Novosoft, we had object-oriented views of both XMI and DAML, so that the details of both specifications were hidden from us.

The IMMACCS build process involves converting XMI into various other forms, as noted above. In order to implement a single framework that can be specialized for different output targets, we created an abstract class, UMLProcessor, which provides methods for using Novosoft to parse an XMI file, along with a set of methods for walking through the resulting objectified version of the XMI.

To use UMLProcessor for the generation of individual target formats, developers can implement a class that extends UMLProcessor, implementing methods that are useful for generating the desired output. In the case of IMMACCS to DAML, we implemented a class named UMLtoDAMLProcessor to form a bridge between the Novosoft representation of the XMI, and an equivalent DAML model.

In general, converting from UML to DAML was straightforward and proved that DAML has more than enough expressive power to represent anything that can appear in a UML class diagram. There was one point, however, that proved to be slightly problematic.

Attribute names in UML can be repeated in different classes. For instance, a tree can have a height and a person can have a height. But in DAML, properties must have unique names. Even if a "height" property is defined in one place to have a domain of person, and another "height" property is defined to have a domain of tree, there is still only one height property. The domain, however, becomes the union of "person" and "tree", which – outside of Tolkien's Middle Earth – is likely to be an empty set. In order to maintain the meaning from UML, we adopted a convention that the name of a property is the name of the class, followed by a period, followed by the name of the property. Thus, the height property of the person class in UML becomes "person.height" in DAML.

Figure 1 presents two sample classes from the IMMACCS domain. Figure 2 displays the result of translating these classes into DAML. In both figures, the information has been

10

simplified by removing extraneous or repetitive details. Other properties of the two classes follow the same pattern as the definition for "IMMACCSObject.dateTime." The concatenation of the class name to the property name eliminates the potential name clash problem discussed in the previous paragraph. All other property names received the same treatment.

```
┌─────────────────────────────────┐
│         IMMACCSObject           │
├─────────────────────────────────┤
│ + source  : string              │
│ + referenceName  : string       │
│ + dateTime  : long              │
├─────────────────────────────────┤
│                                 │
└─────────────────────────────────┘
```

```
┌─────────────────────────────────┐
│            Agent                │
├─────────────────────────────────┤
│ + agentId  : string             │
│ + function  : string            │
│ + activity  : long              │
├─────────────────────────────────┤
│                                 │
└─────────────────────────────────┘
```

Figure 1: IMMACCS classes (example)

```
<daml:Class rdf:about="IMMACCSObject"/>

<rdf:Description rdf:about="IMMACCSObject.dateTime">
     <rdfs:range>http://www.w3.org/2000/10/XMLSchema#long</rdfs:range>
     <rdfs:domain rdf:resource="IMMACCSObject"/>
</rdf:Description>

<daml:Class rdf:about="Agent">
    <daml:subClassOf rdf:resource="IMMACCSObject"/>
</daml:Class>
```
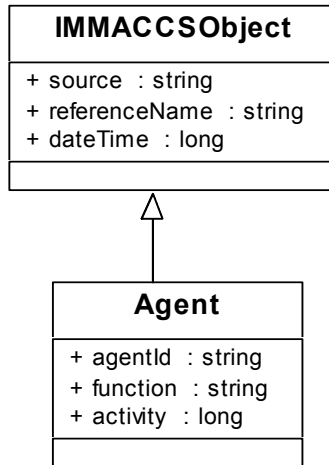
Figure 2: IMMACCS classes translated to DAML (partial)

The IMMACCS to DAML project did not include a user-interface. In the next project, Ontology Merge, we added the ability to view ontologies in tree and table views.

Ontology Merge: As more service providers adopt the Semantic Web and an ontology-based approach to the information they provide, we can expect that there will be a rapid proliferation of ontologies for all domains, include multiple overlapping ontologies for related domains. There will also be a tendency to standardize, since the use of shared ontologies will be extremely beneficial in rapidly producing interoperable systems. However, there will frequently be a need for systems to access and reason about information from multiple ontologies. To do this, we need methods for automatically merging ontologies when necessary.

11

The problem of merging ontologies in general is extremely complicated, and in all likelihood requires some level of human participation to disambiguate terminology. The Future Work section of this document includes a discussion of some of the problems and approaches.

The simplest form of ontology merging, however, can be handled programmatically. This situation occurs when multiple service providers extend the same base ontology in different ways. This is also a reasonable real world scenario: service providers are likely to start with a standard ontology and extend it to suit their own requirements. A consumer whose interest is in information related to a given domain of knowledge is likely to run into sites that have extended standards for that domain in different ways. It would be useful for consumers to have the ability to make inferences using information even when that information deviates slightly from the standard.

For the ontology merge project, we chose to simulate a situation relating to the supply of electricity. In this scenario, we assumed two web-based systems. Once of these systems provides information about the current state of the electrical grid system, while the second is concerned with reporting traffic information. In this case, the information is restricted to reporting on the condition of traffic signals.

We began with a base ontology, which was then extended in different ways by the services, the domains of which involved specializations of the base ontology. All ontologies for this project are included as Appendix A. Our goal for this project was to show that an application could merge information from different services at run-time and successfully reason about instances of classes that were not designed into the application but were discovered as the result of interaction with various services.

In the base ontology, we began with a simple representation for PowerSupply and PoweredObject. Then we constructed relationships between the two:

> PowerSupply suppliesPowerTo PoweredObject .
> PoweredObject powerSuppliedBy PowerSupply .

Next we specialized these classes to create ElectricPowerSupply and ElectricalObject. ElectricPowerSupply is a subclass of PowerSupply, where the "suppliesPowerTo" property is restricted to objects whose class is ElectricalObject, and ElectricalObject is a subclass of PoweredObject, with the restriction that the "powerSuppliedBy" property must be filled by an object whose class is ElectricPowerSupply.

We also included in the base ontology several classes related to drawing simple maps on the screen. Among these classes are Map, Polygon, Point, with associated properties. These classes were mainly included to support part of the planned user interface display functionality, described further below.

We defined a basic ontology for the electric grid. In this model, the class Grid is a subclass of both MapArea and ElectricPowerSupply from the base ontology. This represents the idea that an electrical grid covers a certain geographical area, and supplies electric power. We also developed an experimental set of classes and properties that allow the expression of constraints on one property of an object, based on the value of another property. In this demonstration program, this capability was used to state that the map area associated with an instance of the Grid class will be one color when the grid is

12

working, another color when the grid is not. This assumes a certain type of map display, but does not require it, since each application is free to interpret the display of the map in any way that suits the application's requirements. The semantics of the constraint properties were defined as axioms in the CLIPS expert system shell language and are listed in Appendix B [NASA92].

Having defined concepts relating to power supplies, powered objects, maps, and electrical grids, we next defined an ontology for terminology related to traffic signals. Each traffic signal is associated with an Intersection, which is a subclass of the Location class from the Map ontology. Intersections are associated with two or more Roads, which are named linked lists of Intersections. The TrafficSignal class itself is defined as a subclass of ElectricalObject.

We also defined a set of axioms that implement a subset of the DAML semantics. This allowed us to check validity on various DAML constructs, such as the restriction that the values of a given property must be instances of a specific class. It also enabled us to build a user interface for adding new objects, while making sure that the new objects comply with the semantics of the ontology. If these axioms were to be further extended, it should also be possible to allow the creation of new classes while the application is executing. The validity axioms are again included in Appendix B.

We were not able to design and implement an application that is entirely free of assumptions about the nature of the ontologies that it is working with. Such an application would, in the researchers' opinion, be so generic that it would provide only limited usefulness, perhaps being restricted to simply viewing and modifying classes and instances. We felt it would be more realistic to assume that any application would have *a priori* knowledge of the base ontology, and learn about all extensions. In this case, the application was designed around the base ontology, with its description of PowerSupplies, Maps, and so on. The most important aspect of the application that relies on the base ontology is the map display. We were unable to find any way to implement a program that would be able to infer the best way to display instances of the class Polygon, for example, if we have to assume that the program has no predefined notion of the geometrical concept of a polygon.

The ontology merge application was able to display electrical Grids, in spite of the fact that there was nothing in the application's code concerning the Grid class. The application learned the definition of the Grid class by downloading the ontology from an ElectricalGrid service. When the application learned that Grid was a subclass of MapArea, the application was able to perform the same actions for an instance of Grid that it would for an instance of MapArea. In this way, the application was able to extend its own internal ontology autonomously, without code modifications.

The ontology merge application displays information about classes and instances in a tree control and shows the details in a table. Objects whose classes relate to the Map ontology were also given a graphical display area, where Grid objects are drawn as polygons, roads as connected line segments, and traffic signals as symbolic representations of traffic signals. Since the application was assumed to be aware of the base ontology, it also includes rules that when a power supply is not working, the powered objects that are associated with the power supply will be shut off. The user is able to edit the values of

13

the properties of objects. For instance, the user can change the state of an electrical grid from 'working' to 'not working.' This will cause the traffic signals within that grid to become non-functional as well, even though the Grid ontology has no notion of a traffic signal, the Traffic ontology has no notion of an electrical grid, and the application knows nothing about either one.

The ontology merge project demonstrated that it is possible, under controlled circumstances, to bring together information from various sources, even sources using an ontology that is not previously known to the system using the information. We showed that it is possible to perform inferencing across objects from different ontologies, and to display and use the results. As the Semantic Web becomes more widespread, we expect that this ability will be key to taking advantage of heterogeneous information providers.

Future Directions

During the course of the two years that we have been working on projects relating to the Semantic Web, many useful tools have been created by other organizations. In some cases, these toolsets are the result of large-scale efforts over a period of years. Such tools – a few of which were listed in the Survey of the Field section – exhibit a level of functionality and maturity that make them likely candidates for inclusion in complex dynamic systems. We are looking forward to acquiring and using these tools to provide advanced capabilities for our own systems.

In the course of this work, we have realized both a need and an opportunity for taking the research forward in a related but unexpected direction. It has become apparent that there is a requirement to adapt ontology-related technology to the processing of existing documents, while at the same time building tools that enable the creation of new documents within a semantically aware environment. Along with other tools, this new generation of tools could become the basis of ontology-based knowledge management systems for the intelligent enterprise.

Our future plans are to collect documents from throughout our organization into a single knowledge management system. These documents will be converted from their current format (Word, PDF, LaTEX, HTML, etc.) into a common XML schema (content only, with the presentation markup removed). Using the result as a corpus of instance documents, we will use multiple methods of classification, such as those described in the "Survey of the Field", above, to extract an ontology. The documents will then be marked up by creating OWL instance documents for each one. The semantic markup will allow intelligent search and document assembly, turning the entire store of documents into a warehouse of organizational knowledge. When combined with other task-specific tools such as scheduling, organizational history repositories, and so on, the result will be a system that supports enterprise-wide learning

Conclusions

Defining mappings between ontologies is likely for the foreseeable future to remain the task of humans who are expert in the design of ontologies as well as the domain of knowledge under consideration. When the ontologies of concern are the product of different user communities, there is the extra complication that all mappings must be the

14

product of a process of consensus building. Creation of mappings through completely automated means is unlikely if not impossible.

Our work has shown that in two very controlled and limited cases, it is possible to achieve forms of mapping through the use of defined relational axioms. These types of mapping are useful in themselves, but are insufficient in any wider or more general application. For mapping tasks that do not fall into the categories that we considered, some of the statistical methods or more user-centric approaches listed under Related Work would be more likely to succeed.

In the course of researching possible approaches to this problem, reported methodologies based on the use of statistical modeling to extract or suggest ontologies for an existing corpus of documents have led us to extend our proposed field of investigation in the future. We believe that the same methods that have been show to be of value in producing mappings between documents in known classifications can be used to extract information that will serve as input to a process of constructing ontologies. In turn, these ontologies can become the basis of an enterprise knowledge management system. We look forward to testing this hypothesis in future projects.

Bibliography

[Berners-Lee01] Berners-Lee, T., Hendler, J., Lassila, O. **The Semantic Web** Scientific American, May 2001.

[Booch99] G. Booch, J. Rumbaugh, and I. Jacobson **The Unified Modeling Language User Guide** Addision-Wesley Publishing, 1999.

[daml] www.daml.org

[Doan02] A. Doan, J. Madhavan, P. Domingos, A. Halevy **Learning to Map between Ontologies on the Semantic Web** Eleventh International World Wide Web Conference (WWW'2002), Hawaii, USA

[Dou03] D. Dou, D. McDermott, and P. Qi **Ontology Translation on the Semantic Web**. In *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE2003)*, 3-7 Nov. 2003, Catania, Italy

 [Joachims97] Thorsten Joachims **A Probabilistic Analysis of the Rocchis Algorithm with TFIDF for Text Categorization.** Proceedings of International Conference on Machine Learning (ICML), 1997.

[Maedche02] A. Maedche, B. Motik, N. Silva and R. Volz **MAFRA – An Ontology MApping FRAmework in the context of the Semantic Web**. In *ECAI-Workshop on Knowledge Transformation for the Semantic Web*, Lyon, France, 07 2002.

[McBride03] Brian McBride **An Introduction to RDF and the Jena RDF API** http://jena.sourceforge.net/tutorial/RDF_API/index.html, August 2003

[McGuinness00] D. McGuiness, R. Fikes, J. Rice, S. Wilder **An Environment for Merging and Testing Large Ontologies**, in *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, Breckenridge, Colorado, USA, 12-15 April, 2000

[McGuinness03] Deborah L. McGuiness and Frank van Harmelen, eds. **OWL Web Ontology Language Overview.** World Wide Web Consortium, Web Ontology Language Working Group. August, 2003. http://www.w3.org/TR/owl-features/

[Pease00] A. Pease, V. Chaudri, F. Lehmann, and A. Farquhar **Practical Knowledge Representation and the DARPA High Performance Knowledge Bases Project**, in *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000)*. San Francisco, CA, 2000

[Oullet02] Roxane Ouellet and Uche Ogbuji **Introduction to DAML**
http://www.xml.com/pub/a/2002/01/30/daml1.html, January 2002

[Pekar02] Viktor Pekar and Steffen Staab **Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision.** Proceedings of the Nineteenth Conference on Computational Linguistics, 2002

[Pohl03] Jens Pohl, Anthony Wood, Kym Pohl, "IMMACCS: An Experimental Multi-Agent C2 System". 5th International Command And Control Research and Technology Symposium. Canberra, Australia. October 24-26, 2000.

[NASA92] Software Technology Branch, **CLIPS Reference Manual.** Johnson Space Center, Houston, Texas, 1992.

[Su03] X. Su, S. Hakkarainen and T. Brasethvik **Improving semantic interoperability through analysis of model extension.** Proceedings of the CAiSE '03 Doctoral Consortium. Velden, Austria, 2003.

[Wache01] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner **Ontology-Based Integration of Information - A Survey of Existing Approaches** Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing,

Appendices Available Upon Request

16

Design Methodologies for Analog/Mixed Signal VLSI Systems Applied to Infrared Focal Plane Arrays

William Ahlgren
Anton Clarkson
Electrical Engineering Department

Joseph Grimes
Computer Science Department

Cal Poly State University

DESIGN METHODOLOGIES FOR ANALOG/MIXED SIGNAL VLSI SYSTEMS
APPLIED TO INFRARED FOCAL PLANE ARRAYS

ONR GRANT #N000-14-02-1-0754, PROJECT #55371

Report for the period July-September 2003

by

William Ahlgren and Anton Clarkson
Electrical Engineering Department

and

Joseph Grimes
Computer Science Department

Abstract
We are developing a design methodology for analog/mixed signal systems, based on
emerging CAD tools. Our example system is a high-performance image sensor for
infrared wavelengths, a so-called infrared focal plane array. We are developing a pixel
design and sensor architecture to exploit the opportunities presented by the continuing
advance of CMOS technology, following Moore's Law. Current image sensors are
limited to on the order of 10 transistors per pixel, whereas new technology will enable
thousands of transistors per pixel, thus enabling revolutionary new approaches to pixel
design and sensor architecture. To exploit this opportunity, we are developing a design
that implements pixel-level analog-to-digital conversion combined with a novel readout
architecture. We are also exploring the implications of System-on-Chip integration,
especially issues of data communication between image processing sub-systems built on
the same chip with sensor array. This design will be applicable in many systems
requiring image sensors, and especially systems requiring very-long-wavelength infrared
and terahertz sensors. These sensors are used in diverse applications, including ballistic
missile defence, concealed weapons screening, medical imaging, and astronomy.

Summary of Previous Work
Following is a summary of previous work performed on this project. Please refer to our
past reports for more details.

Computational Infrastructure Through partnership with Cadence Design Systems, Cal
Poly has access to state-of-the art analog/mixed signal integrated circuit/system
computer-aided design tools. We have developed at Cal Poly a hardware platform for
these tools based on Sun enterprise servers, networked workstations, and 26 SunRay
smart terminals for teaching large classes. The facility provides 24-hour access for both
faculty and students doing project work. We have developed and documented
installation and maintenance procedures for the tool suite provided by Cadence, and we

are continuing to develop training materials and publicize the availability of these tools for use by faculty and students.

State-of-the-Art-Review We prepared a review of the current state-of-the-art for CMOS image sensors and infrared focal plane arrays. This review confirmed that there is a need to advance the current technology through full exploitation of the opportunities made available by the continuing advance of CMOS technology following Moore's Law. This makes possible the incorporation of ever-more signal processing power within a single pixel. Pixel sizes are fixed by optical considerations (Rayleigh criterion), and are quite large for infrared wavelengths. To exploit the smaller feature sizes made available in new technology generations, more transistors should be incorporated into each pixel. How best to use these transistors is the issue we address. Our design implements pixel-level analog-to-digital conversion, thus following the trend of all modern mixed-signal technologies to push this critical signal processing step as toward the front-end of the processing chain as possible.

Pixel Design and Sensor Architecture The next CMOS technology generation, with 90 nm minimum feature sizes, will enable 512 transistors per 20 μm pixel, and 2048 transistors per 40 μm pixel. This technology will be widely available in the next few years. Moore's Law will not stop there. Intel has already announced plans for a 45 nm technology by 2007, which would quadruple the number of transistors per pixel. The availability of thousands of transistors per pixel, rather than the current 4-12, suggests that revolutionary approaches to pixel design are needed. We elected to develop one of the simplest of these, as follows: the current-source output of the photosensor is first converted to frequency, and then fed to a digital counter. That is, the current source is first converted to a sequence of pulses, the number in a given time interval proportional to the light intensity. By counting these pulses, a digital representation of the light intensity is obtained. Two key components were identified for initial development: the current-to-frequency converter and the counter.

Progress Since Last Report

During Summer 2003 we continued work on pixel design and sensor architecture, and also initiated a state-of-the art survey to identify issues in on-chip data communications for SoC (System-on-Chip) sensor/processor architectures.

Sensor Architecture In our previous report we identified the most important parameters characterizing an image sensor:

N, the number of columns (x-coordinates) in the pixel array

M, the number of rows (y-coordinates)

K, the number of different spectral response detectors (colors) per pixel

S, the number of sensitivity bits in the digitized pixel output

R, the number of resolution bits

Q, equal to $R + S$, the total number of quantization bits

P, the number of parallel lines on which the sensor output data stream appears

Ti, the integration time for image acquisition

Tf, the time from frame-start to next frame-start

Tr, the time required to readout a frame of data
NMK is the number of analog data values per frame. Each of these values may be
obtained at different sensitivities (light-to-electrical conversion factors), specified by S
sensitivity bits, and is converted to digital form with R resolution bits, resulting in Q = R
+ S quantization bits. The output of the sensor is on P parallel lines. The integration (or
exposure) time is Ti and frame time is Tf. In conventional expose/readout mode, Tr =
Tf-Ti is the time available for readout of all the values before a new frame begins. Data
acquisition can be pipelined if simultaneity within a given frame is not mandatory,
resulting in Tr almost equal to Tf, if needed. Alternatively (or additionally), increasing
the number of parallel outputs P will reduce data rate. This is important because power
consumption is proportional to data rate, so strategies for keeping the data rate low
should be considered, even if not essential to meet other design objectives.

For initial development we have selected a sensor with single-color operation (K = 1) in
which each of the M rows is selected in sequence, and all N columns are read in parallel
into a N x Q memory cache (an output buffer "shift register" having Q bits in parallel
rather than just one). The data is then be multiplexed out on P parallel lines. Figure 1
shows a representation of a sensor architecture with M = 2 and N = 3, and Q = R = P = 4.
A full-sized sensor might have M and N on the order of 1000, and R in the range 8-16.
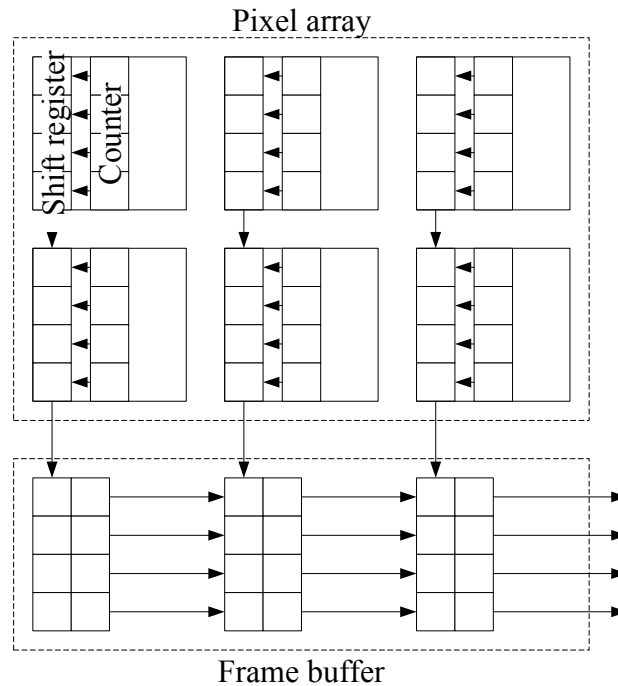
Figure 1. Sensor architecture with separate counter and shift-register in each pixel.

In the architecture shown in Figure 1, processing elements in each pixel (not shown in the
figure, but occupying the blank space on the right-hand side of each pixel), feed a pulse
train to the counter. The frequency of the pulse train is proportional to the intensity of
the radiation incident on the pixel, thus the value stored in the counter after a fixed

integration time is a digitized representation of the signal. At the end of the integration time, the contents of the counters are shifted horizontally into the shift-registers. The shift registers of all pixels in a column are chained together vertically to enable readout in a manner equivalent to that of a CCD image sensor. The contents of the entire frame is strobed into the frame buffer memory during the readout time, and the pixel counters are cleared prior to initiating another integration time.

Because the counter and shift register in each pixel do not operate simultaneously, space can be saved if the two are replaced with a single shift-register counter. This is a device that has two states, shift-register and counter, and which state it is in can be selected by a control signal. One such device (the only one we are aware of) is a linear feedback shift register (LFSR). When operating in counting mode, this device cycles through a series of bit patterns that require decoding in order to associate them with the corresponding binary numbers (this will be explained in more detail below). Consequently, a decoder block is needed before further processing is performed on the resulting video stream. The architecture corresponding to this approach is shown in Figure 2.
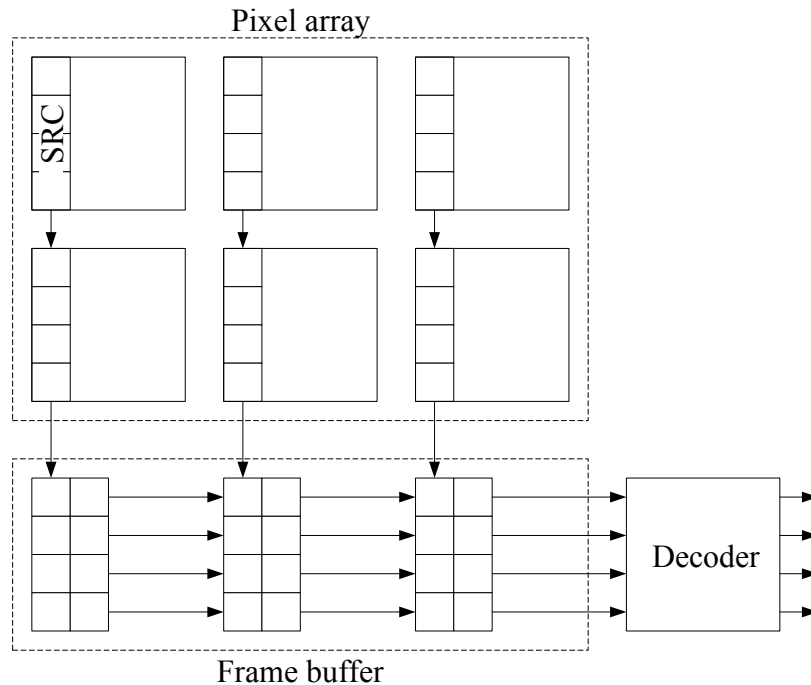
Figure 2. Sensor architecture with a single shift-register-counter in each pixel.

The approach shown in Figure 2 has two disadvantages. First, the additional decoder block takes up space on the chip. Second, this architecture will consume more power. The latter is the more serious issue, and has two components, the additional power required to run the decoder, and the additional power consumed by the counter in each pixel if a Gray code cannot be implemented.

Despite these drawbacks, we believe the architecture of Figure 2 merits further investigation, which we have initiated in this report period. We expect in future to compare the overall performance of both architectures.

Pixel Design Two key elements of the pixel design that were previously identified are the current-to-frequency converter and the counter. During this report period we have focused on the design of the counter. Although a large number of transistors are potentially available for each pixel, we are still interested in minimizing the number of transistors used per component, in order to maximize the applicability of the design to a wide range of wavelengths and fabrication technologies. One way to minimize transistor count is to use a shift-register-counter (SRC) rather than a simple counter. During the integration time, the SRC functions as a counter; during readout, the SRC functions as a shift register.

SRC Design A linear-feedback shift register (LFSR) is a way to build an SRC, with minimal a number of components. This SRC design is well-adapted to imaging applications (Refs. 1 and 2). Figure 3 shows a four-bit SRC configured to be incorporated into a pixel. The components shown are D flip-flops (DFFs), an XNOR gate, and 2:1 multiplexers (MUXs), which are used as SPDT switches.
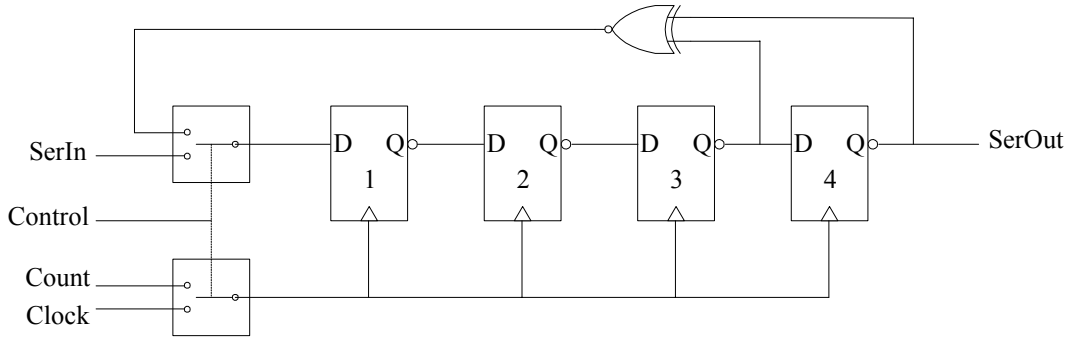
Figure 3: Shift-Register-Counter based on LFSR principle.

When the MUXs are set to count mode (switches in upper position), the SRC functions as a counter. The "Count" input is the pulse train generated by the current-to-frequency converter in the pixel. When set to shift mode (switches in lower position), the SRC functions as a shift register. The "Serial In" input is the data coming from the next pixel up in the column; the "Serial Out" output goes to the next pixel down in the column.

The DFFs are inverting; at the clock edge, $\overline{D} \rightarrow Q$. The output of the XNOR is high when both inputs are the same, i.e. $D1 = \overline{Q3 \oplus Q4}$. Using these relations, the following table is generated:

Table 1: States of the SRC in Counter mode.

Pulse#	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Q1	0	1	1	1	0	1	1	0	0	1	0	1	0	0	0	0

Q2	1	1	0	0	0	1	0	0	1	1	0	1	0	1	1	1
Q3	0	0	0	1	1	1	0	1	1	0	0	1	0	1	0	0
Q4	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	1
D1	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1	0

The pattern 0101 is chosen as the initial pattern because it is easy to set it by tying SerIn high and pulsing Clock four times. Note that the pattern 1010 (which would be created by setting SerIn low and pulsing Clcok four times) does not appear; it is the lock-up pattern associated with this LFSR. In general, for maximal-length LFSRs such as this one, there are 2^N-1 patterns generated, not 2^N, and the last pattern does not occur (and if did occur, the counter would lock-up in that state). The device can count from 0 to 14, generating the 15 numbers in coded form. 0000 is coded as 0101, 0001 is coded as 1101, 0010 is coded as 1001, etc. Since pulse counts 0000 and 1111 are both coded as 0101, it must be assumed that 1111 will never occur, or else other means have to be provided to deal with that possibility. A Gray code counter, in which only one bit changes at a time, is the ideal for our application because it minimizes energy dissipation. The pseudorandom code generated by the LFSR is not optimal in this respect. Further, a decoder will have to be provided further down the signal processing path to the binary or some other more meaningful code. Nonetheless, this SRC design is attractive for its simplicity, leading to a low transistor count and minimal area realization.

DFF Design The SRC shown in Figure 3 uses inverting FFs and gates because this helps to minimize the transistor count in static CMOS realizations of the circuit. We prefer to create a baseline design using static logic in a counter function, since at long integration times and low signal levels, stored values may have to remain valid for relative long times, of the order of milliseconds. Static storage elements avoid the complexity of refresh logic. In future, the relative merits of using dynamic processing elements will need to be further studied, especially when evaluating the architecture of Figure 1. To minimize transistor count with a static design, we use a 4-transistor latch as described in Reference 3. To function as a FF, the latch is controlled by a pulsed clock. The circuit is shown in Figure 4.

Simulation of this circuit was performed using the Cadence SPECTRE circuit simulator, with transistor models from an example library. The simulation showed that a four-bit shift-register comprised of these cells can operate correctly over a limited range of clock pulse period and duty cycle. This study will be continued with better-defined transistor models, corresponding to a current or extrapolated DSM technology, and will be extended to a layout-based model including the effect of interconnect parasitics, and also including a clock pulse generation circuit.
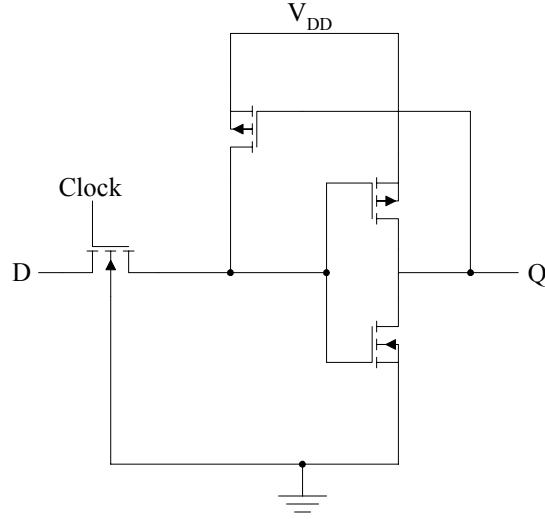
Figure 4: Inverting DFF design, a four-transistor latch with pulsed clock. Note that $\overline{D} \rightarrow Q$ while clock is high.

XNOR and MUX Design In realizing XNOR and MUX functions, we need not be as concerned to minimize transistor count, since relatively few of these components are needed per pixel. Still, we are exploring circuits that are as lean as possible. Conventional static XNOR designs use 8 transistors, and a well-known variant uses 6. We are evaluating a 4-transistor design described in Reference 4, and shown in Figure 5.
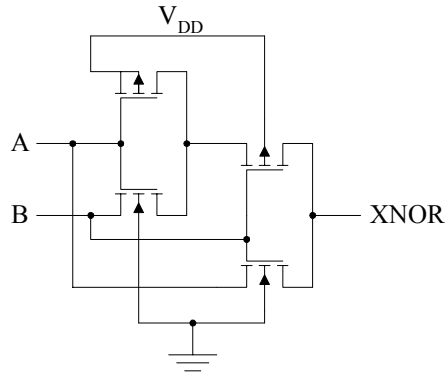
Figure 5: Four-transitor XNOR design.

A conventional 2:1 MUX, which can implement one of the switches shown on the left side of Figure 3, requires 8 transistors (two transmission gates and two inverters); thus 16 transistors would be required to implement the double switch required. While perhaps acceptable, further study, perhaps along in the direction indicated by Ref. 5, can be expected to result in a more compact realization of this function.

Intra-Chip Communication Networks provide a ubiquitous communication service analogous to that of water, transportation, and electrical. Because of their flexibility, they are much closer in this way to the transportation services. In the past, networks have provided an interchange of information between end systems such as computers, printers, disk drives, etc. Because of the density and size of modern chips, large components such as multiple computer processors, memory, and other computer components are being fabricated on a single chip. Because of the need to interchange large quantities of information between these components, the need for intra-chip networks has developed. In project, it is anticipated that large components such as sensor, digital signal processors, control, and memory will be present on a single chip and it will be necessary to have an efficient high-speed communication interface between them.

The approach to be used would be the use of threaded processing between multiple digital signal processors combined with instruction level parallelism from a single instruction stream. It is anticipated that it would be best to have local memory associated with each of the processors. It is believed that a message passing programming model and a networked communication for message exchange would provide greater scalability and efficiency than the bus approach of the past. In order to achieve the desired goals, it would be necessary to overcome the inter-component communication bandwidth limitations, component synchronization issues, dealings with the latency issues of the network, and difficulties of existing parallel computing models. The software developers have had responsibility for these issues in the past and must be eliminated to some degree through innovative parallel design architectures and more advanced parallel programming strategies that use as fully as possible both instruction level programming and thread level programming.

An important step toward achieving this goal will be the development of efficient methodologies for simulating systems on a chip model. A necessary element of this approach will be the development of artificial intelligent techniques using expert knowledge to direct the schedule of activities to be simulated. This will be achieved through the artificial intelligent techniques driving the scheduling of the sequence of software execution as well as hardware components.

This project provided an opportunity to research the various strategies that have been developed (References 6 to 12), and do some initial consideration of how communication networks might be used in the implementation that is being considered in this project. Significant additional work must be done to extend the prior work to this particular application and then develop the special simulations that must occur in order to determine most efficient strategies.

Further Work
The next task is to continue the detailed pixel design. This task entails additional transistor-level simulation of the circuits discussed above, as well as alternative designs. In addition to the shift-register-counter, the current-to-frequency converter and pixel-level control circuits must be designed and simulated. Physical layout and verification must be performed. We have more to work to do to improve our design infrastructure.

We need to install and gain experience with process design kits that could be used with MOSIS fabrication services. We also need to begin to explore the creation of our own process design kit, aimed at a notional advanced fabrication process. For example, we might choose a 45 nm process, projected for availability in 2007.

Next, we need to design other sensor-level sub-systems including control, clock generation, row select, frame buffer, decoding, and multiplexing functions.

We also would like to explore system-level simulation of the proposed chip architecture using the Verilog-AMS hardware description language. A key aspect of system-level design is the data communications between functional modules, for example, between the sensor array and on-chip DSP blocks. A network-on-chip approach to data communications will be explored. Noise modeling and simulation will be incorporated, including noise due to fundamental processes, and noise due to unintentional signals. Design for manufacturability, including design for test, will be emphasized. We will assess our design using Monte Carlo methods to simulate the random variation during manufacturing of key parameters, and their effect on yield. Test is a major cost driver, to be addressed by creating software test benches for our design. A software test bench is an HDL program used to verify ("test") another HDL program. This is a standard design technique for digital VLSI systems, which we expect to be able to extend to analog/mixed signal systems.

The final task is to integrate the subsystems into the complete SoC, and perform chip-level physical verification (parasitic extraction, simulation, signal integrity analysis).

References
1. P. Fischer, "An area efficient 128 channel counter chip." *Nuc. Instr. & Meth. in Phys. Res. A*, <u>378</u>, 297-300 (1996).
2. J. Hausmann, *Entwicklung und Test einer integrierten Signalverarbeitungselektonik für bildgebende Pixeldetektoren*. Unpublished MS thesis, Universität Bonn.
3. Qiu Xiaohai and Chen Hongyi, "Discussion on the low-power CMOS latches and flip-flops." *Proc. 5th International Conf. on Solid-State and Integrated Circuit Technology*, IEEE, 1998, pp. 477-480.
4. H. T. Bui, A. K. Al-Sheraidah, and Y. Wang, "New 4-transistor XOR and XNOR designs." *Proceedings of the 2nd IEEE Asia-Pacific Conference on ASICs*, IEEE, 2000, pp. 25-28.
5. K. Yano, Y. Sasaki, K. Rikino, and K. Seki, "Top-down pass-transistor logic design." *IEEE J. Solid-State Circuits* <u>31</u>, 792-803 (1996).
6. Benini, L. and DeMicheli, G., "Networks on Chips: A New SoC Paradigm," *IEEE Computer*, Jan. 2002, pp. 70 – 78.
7. Forsell, M., "A Scalable High-Performance Computing Solution for Networks on Chips," *IEEE Micro*, September-October 2002, pp. 46 – 55.
8. Guerrier, P. and Greinier, A., " A Generic Architecture for ON-Chip Packet-Switched Interconnections," *Proc. Design, Automation, and Test in Europe (DATE)*, IEEE CS Press Los Alamitos, Calif., 2000, pp. 250 – 256.

9. Leon-G., Alberto and Widjaja, I., *Communication Networks: Fundamental Concepts and Key Architectures*, Second Edition, McGraw Hill, 2003

10. Perlman, R., *Interconnections: Bridges, Routers, Switches, and Internet Protocols*, Addison- Wesley, Reading, Massachusetts, 2000.

11. Richter, Kai, Jersak, Marek, and Ernst, Rolf, "A Formal Approach to MpSoC Performance Verification," *IEEE Computer*, April 2003, 60 – 67.

12. RFC 3000, J. Reynolds, R. Braden, S. Ginoza, and L. Shiota, eds., *Internet Official Protocol Standards*, November 2001

Optical Detection of Micron-Sized Airborne Particulate Matter

Thomas Bensky

Physics Department
Cal Poly State University

Final Report:

California Central Coast Research Partnership (C^{3}RP)
Optical Detection of Micron-Sized Airborne Particulate Matter

November 10, 2003

Submitted by:
Dr. Thomas Bensky
Dept. of Physics
Cal Poly State University
San Luis Obispo, CA 93407
email: tbensky@calpoly.edu
Tel: 805-756-7655

1

Contents

1 Project Introduction and Summary

The underlying goal of this project was to see if laser light could be used to detect the presence of particles that have micrometer (μm, millionths of a meter) size dimensions. It is hoped the results of this work could contribute to an understanding of how we might detect airborne hazards, such as the airborne Anthrax spores, which is known to have a size of approximately 100μ m. The original idea for this goal came during the "Anthrax in the mail" scare in late 2001/early 2002. We began wondering if something the size of Anthrax (100 μm) could be optically detected (e.g. using a laser beam). During the term of this research grant, we have made excellent progress toward these goals.

Although airborne micron-sized particles are invisible to the naked eye, and pass right through common air-filtration systems, the premise of this work is that laser light would somehow interact with small (micron-sized) particles (μparticles), giving us an indication of their presence and size. Indeed this happens. We have found that as laser light penetrates a sample of test μparticles, most of the light passing right through. *Some of the light*, however, scatters off of the μparticles, sending the light into random directions with respect to the incoming laser beam. The scattered light levels are very low, much less than a conventional photodiode-detector could respond to. For this reason, we purchased a Hammamatsu photon-detector, which is able to respond to the arrival of *individual photons*, which are the smallest quanta of light. So in short, this work detects μparticles based on the the presence of a single photon that scattered off of the particle. In the actual experiments, we placed the detector at 90° with respect to the laser beam. This creates a maximal discrimination between photons in the original laser beam, and those that could leave the beam by scattering off of a μparticle.

In the experimental setup, the photon-detector is connected to a 500 MHz digital oscilloscope. A sample trace taken with a test sample of 10μm particles in place is shown here.
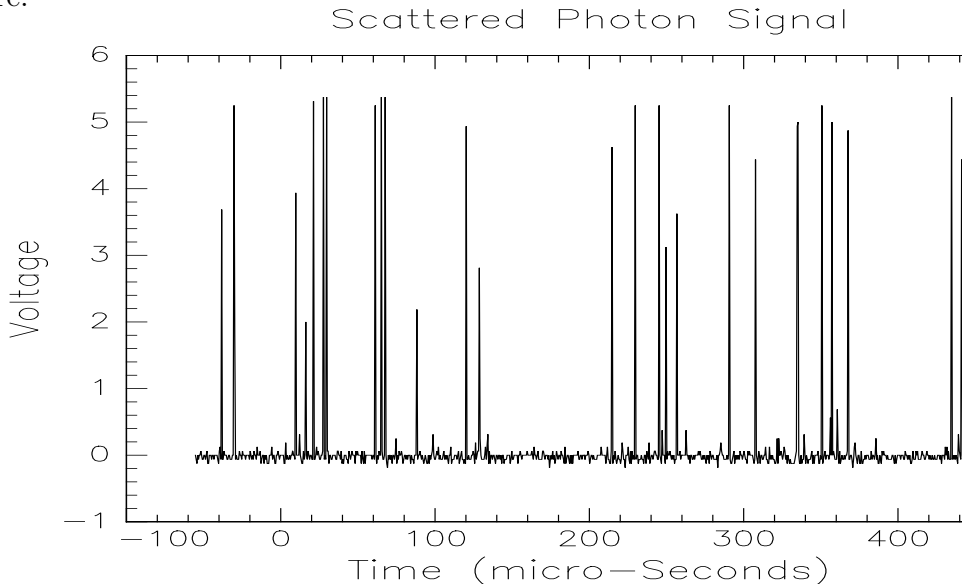
In this report, this raw signal is called the "scattered photon signal." Each peak represents the arrival of a photon, that was scattered by a μparticle, out of the initial laser beam. The horizontal axis corresponds to the arrival time of the photon, and the vertical axis is in voltage (the peak height produced by the photon detector).

3

So, our indication that a μparticle is present is at hand. The scattered photon signal. Processing this signal will actually reveal the size of the particles that generated this signal. Only a small fraction of the peaks in the above figure exist when the sample is removed, are dubbed "background" photons.

2 Supplies Purchased

We had almost no equipment needed to implement the ideas outlined in the original grant proposal. Here is a summary of the items purchased.

Hammamatsu Photon Detector. For detecting the arrival of individual photons.

Optical Table. A 4 foot by 6 foot optical table was purchased to house the experiment, and allow for ease of experimental configuration. It has been installed in the lab space of Building 52, Room C22.

Particle Samples. Our μparticle samples were 1, 5, 10, and 100μm polystyrene spheres. All of these samples are similar in size to Anthrax spores. These spheres were dissolved in distilled water, and placed in a 0.5cm x 0.5cm x 2 cm cuvette.

Optics Hardware. Support hardware was needed to mount the photon detector, laser, optics, and other components.

Diode lasers. These served as our core light source.

Computer and Interface Electronics. A computer was purchased, as was GPIB interfacing equipment. This allowed the computer to acquire data from the oscilloscope (see the photon Figure above), process the data, and compute the particle size.

Salary. T. Bensky (the P.I.) was paid for approximately 100 hours of work on this project (to date, much more than 100 hours has been invested in this work, however.)

Future Possibilities. A few components were purchased to investigate extensions of this work. These ideas were not included in the initial proposal, and were ideas that arose as the work progressed. In particular, we wish to run this experiment with the laser in pulsed mode (not continuous-wave mode). To modulate our diode laser, a Kerr cell and nanosecond pulse generator was purchased.

3 Experimental Procedure

To begin, we used existing equipment to drive a 30 mW 632 nm diode laser. This laser system had sufficient light intensity for this work. We also have limited control over the wavelength, which we've temperature tuned between approximately 625 and 640 nm. For our test samples, we acquired 1.0, 5.0, and 10.0 μm sized polystyrene spheres. The spheres are dissolved in a small cuvette filled with water. The laser beam is aimed at the cuvette,

which subsequently penetrates the water/sphere sample. The basic experimental setup is shown below.

Cuvettecontaining Micro-spheres
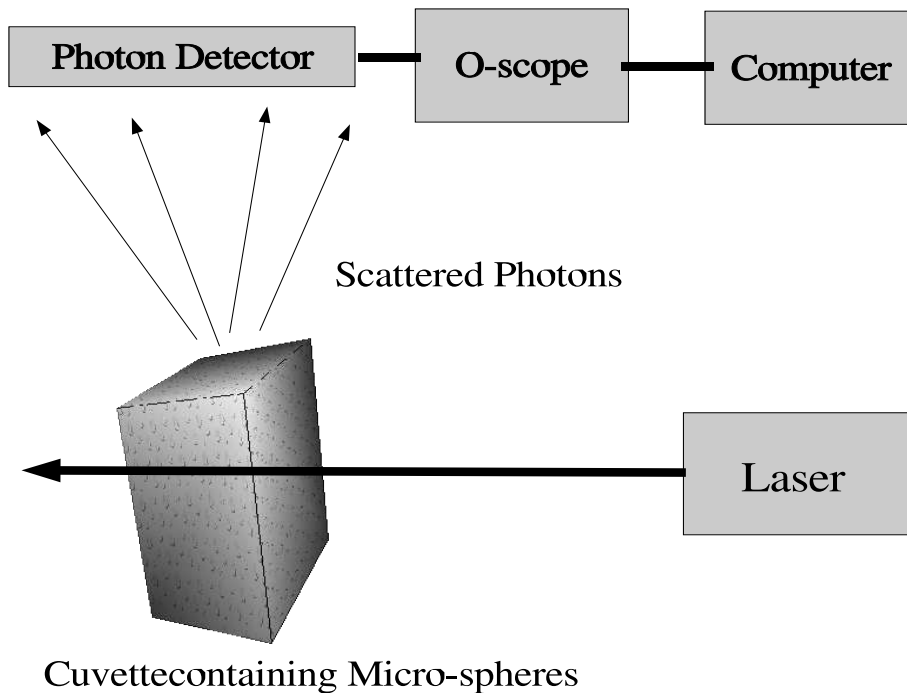
The nature of the signal observed on the detector is as follows. Particles with μm sizes are constantly bombarded by the molecules in the solvent (water in this case). Airborne Anthrax is bombarded the the surrounding air, which can keep it airborne for up to 2 weeks. These bombardments constantly keep the particles moving about, in Brownian motion. Their movement and positions are related to their size, and will dictate whether or not they will move in and out of the laser beam, scattering photons out of it, and into the detector. For a cuvette, containing a given particle size, the scattered photon signal on the oscilloscope is our raw data. Hidden in this data are clues as to the size of the particles in the cuvette.

4 Analysis Procedure

Careful inspection of the scattered photon signal (above), reveals that the photon arrival times are correlated. That is, they seem to arrive in "clumps" and are not randomly distributed in time. The correlation time of this signal, τ_c is somewhat like the "average clumping time" of these photons. In other words, on average, how long in time does a typical clumping of photons last?

A measure of τ_c can be found using the correlation function, which is defined as

$$Corr(\tau) = \int_{-\infty}^{\infty} I(t)I(t+\tau)dt, \qquad (1)$$

where $I(t)$ is the photon count function, which here is the raw oscilloscope "scattered photon signal." There is a simple way of understanding what the correlation function mea-

5

sures. In short, $I(t)$ is multiplied by itself with a time lead or lag (τ), $I(t+\tau)$. The product is then integrated. The more overlap (or similarity) that exists between the $I(t)$ and $I(t+\tau)$, the larger the value of the integral will be. Hence, the τ_c is a measure of how self-similar $I(t)$ is in time. For the scattered photon signal, this will be a measure of the photon clumping time. For a typical photon signal, software we wrote computes a correlation function that looks like this:
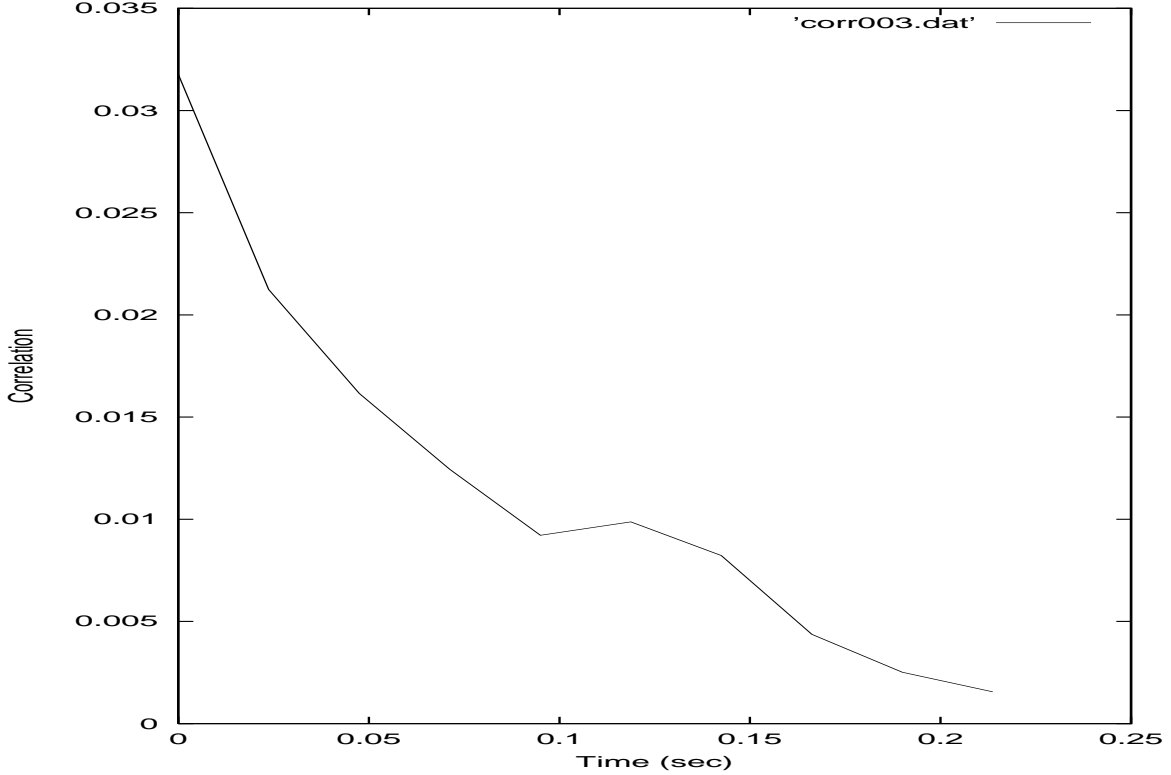
The parameter τ in Equation 1 is plotted on the x-axis, and the result of the integral on the y-axis. Clearly the scattered photon signal becomes less and less similar to itself for larger and larger lead/lag times, τ. For this experiment, the correlation is exponential, following the function

$$Corr(\tau) = Ae^{-B\tau}. \tag{2}$$

Our strategy then is to acquire a scattered photon signal, calculate the correlation for it, fit an exponential to the correlation, then use the "B" parameter in Equation 2 to compute the particle size. The resulting equation for the particle size (radius r) can be found from

$$r = \frac{2kTK^2}{6\pi\eta B}. \tag{3}$$

This theory was not developed by us, and can be found in the literature (in particular, F. Reif, *Fundamentals of statistical and thermal physics*). It is related to how particles move in three-dimensional Brownian motion. The parameters here as as follows. k is the Boltzmann Constant ($k = 1.38 \times 10^{-23}$ J/K), T is the temperature ($T = 316$ K), K is the scattered photon wavelength ($\frac{2\pi}{\lambda} = 1.87 \times 10^7$ m^{-1} at $\lambda = 630$ nm), and η is the viscosity of water, which is $\eta = 936 \times 10^{-6}$ Poises.

6

5 Experimental Results

We ran the experiment, as described above, for three different sizes (1, 5, and 10 μm) of micro-spheres, dissolved in distilled water. For each size, we took several runs and computed the particle size. We histogrammed the results, which are presented here.
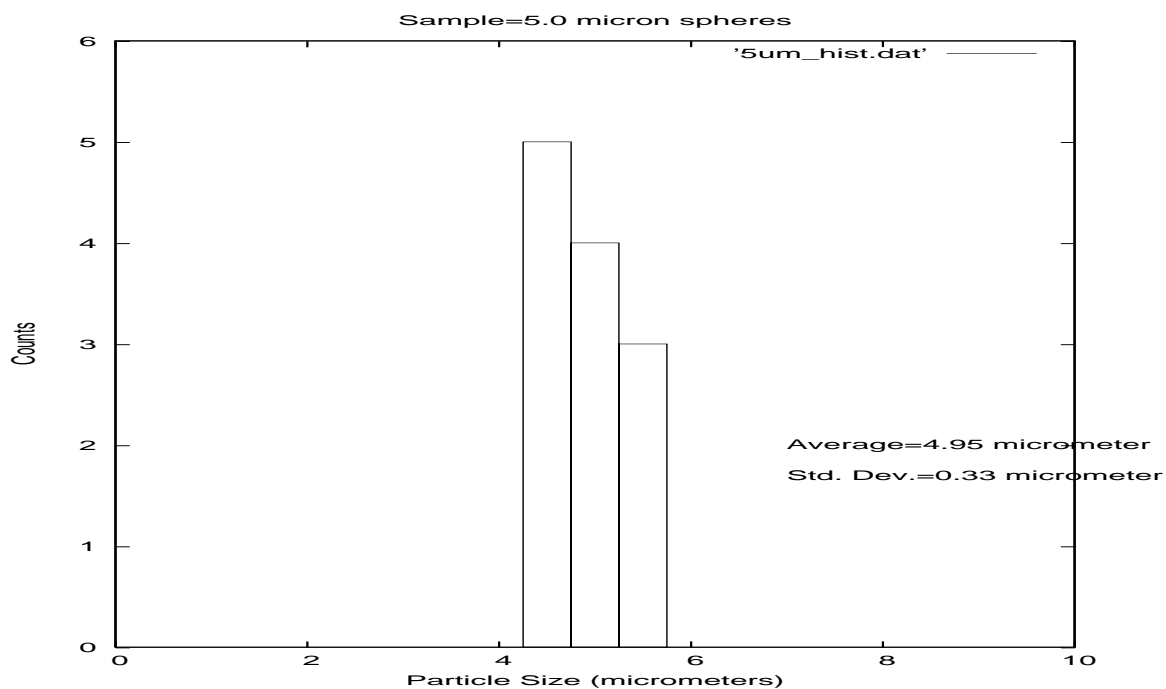
5.1 1 μm spheres

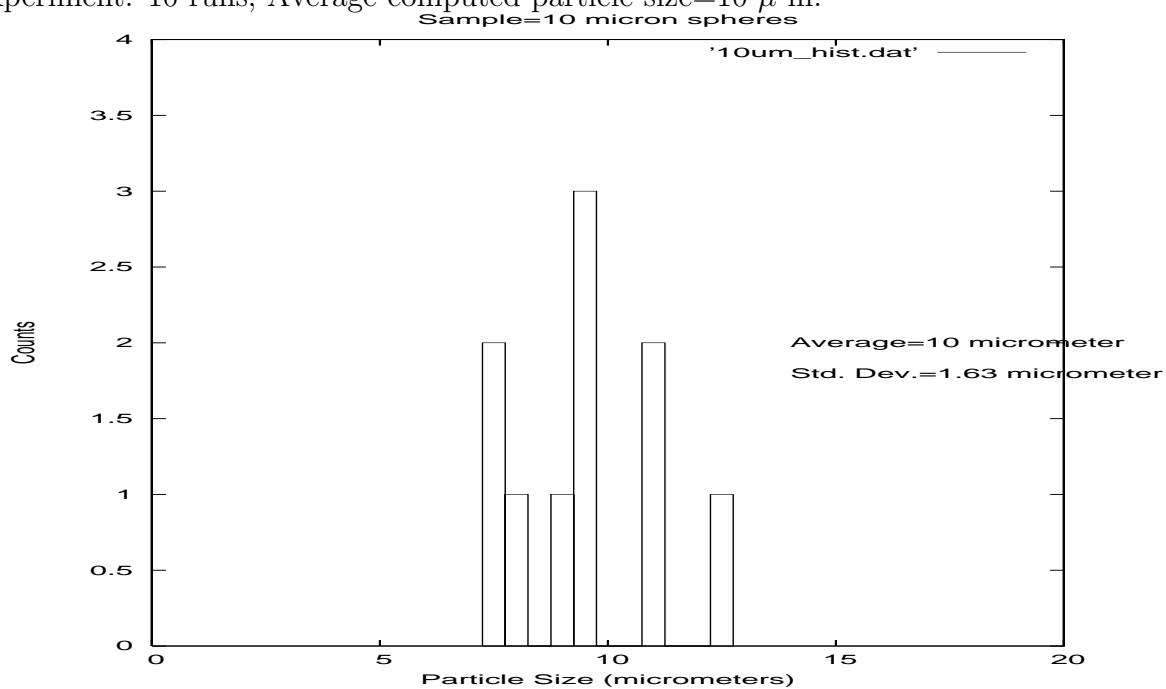Experiment: 20 runs, Average computed particle size=0.84 μ m.
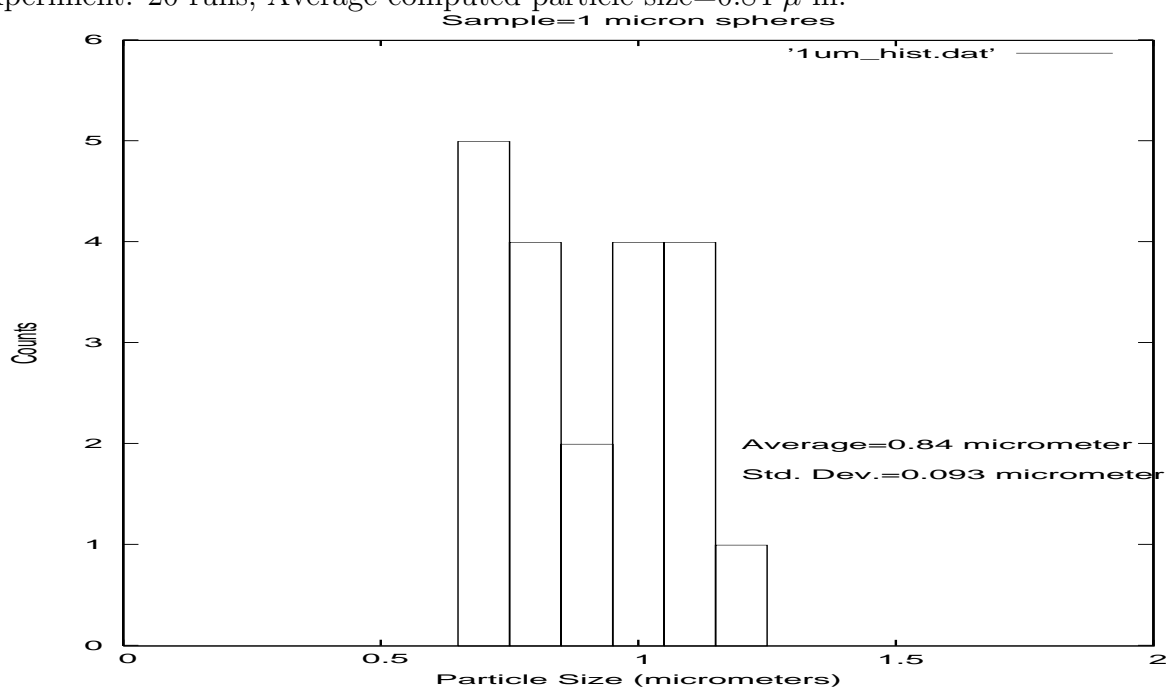
5.2 5 μm spheres

Experiment: 12 runs, Average computed particle size=4.95 μ m.

7

Sample=5.0 micron spheres

Average=4.95 micrometer

Std. Dev.=0.33 micrometer

5.3 10 μm spheres

Experiment: 10 runs, Average computed particle size=10 μ m.

Sample=10 micron spheres

Average=10 micrometer

Std. Dev.=1.63 micrometer

In all cases, performed over several days, we see a clear tendency of our data to fall near and around the expected particle size.

8

6 Conclusions

The original premise of this grant proposal was to see if particles with μm sizes, in a defined volume could be discerned using a laser beam. We feel as if this goal has been achieved. The primary shortcomings of our conclusions, as of this writing, is that this technique has not been tried on *airborne* particles. The Anthrax virus, at least what troubles the United States Postal Service, is airborne. At this point, there are two outstanding questions in adapting this work to the airborne case:

1. A suitable sample. We need to find and acquire a sample of μparticles that could remain airborne in a defined volume. Our best guess it to construct a small chamber with a circulating fan to "blow" the particles around.

2. Theory. The theory used in this work was for particles executing Brownian Motion. It is unclear whether this theory would hold with persistant air current.

We believe the technique presented here would be adaptable to an airborne particles. The outstanding issue outside of the two just mentioned are in *particle concentration*. In our test here, we could increase our particle concentration, as needed, to achieve a suitable signal to noise ratio at our detector. From what we've learned, Anthrax spores "float" through the air *by themselves*, essentially as a single, micron-sized particle in a given volume (like a room). Certainly a single spore will not produce enough scattered photons to build up a suitable data set.

Despite this low concentration, we still believe this technique has potential for detecting airbone micron-sized particles. How? The low concentration is a problem for this reason. The particle to be detected must interact with the laser beam (many times), so photons may be scattered off of it, and a signal at the photon detector may be registered (or, many similar sized particles must interact with the laser beam). A small particle "floating" near the top of a room may not interact (or take along time to interact) with a laser beam situated near the floor. And one interaction isn't enough. Many interactions are needed to produce a scattered photon signal that can be analyzed.

However, this problem might be overcome with an *intense, scanning* laser beam. A volume may be defined, perhaps a 1 m^3 volume (a mail scanner might have such a dimension). Simple optical elements could resize the laser beam to be 1 m in size. Such beam expansion would greatly reduce the laser intensity, but a more intense (10's of Watts) laser could be used. Multiple passes of the laser scanner through the volume could produce multiple scattered photons even off of only a few particles.

9

Non-Linear Aircraft Landing Gear Simulation

Daniel J. Biezad
Stephen Kubic
Grant Kobayashi

Aerospace Engineering Department
Cal Poly State University

Non-Linear Aircraft Landing Gear Simulation

by

Daniel J. Biezad, Ph.D.
Principal Investigator

and

Stephen Kubic and Grant Kobayashi
Students, Aerospace Engineering Dept

November 1, 2003

AEROSPACE ENGINEERING DEPARTMENT
California Polytechnic State University
San Luis Obispo

TABLE OF CONTENTS

LIST OF FIGURES

1

ABSTRACT

In attempt to further de-linearize the six degrees of freedom model currently used by the Calpoly Flight Simulator, gear drag forces were added in order to replicate its effects on the handling qualities of actual aircraft. Modifications to both the C++ coded S-function and alterations to the counterpart gear configuration text file were accomplished to complete the assigned task. A Simulink model was created to handle the newly-assigned task of gradual lowering and rising of the gear devices. The result is both a gear configuration that creates the necessary pitching moments induced when lowered, and code to account for ground, as well as in-flight, force production. Lastly, effort was put forth to develop a data read system from Excel to efficiently aid in rapid prototyping using the flight simulator text files.

NOMENCLATURE

SYMBOL	DEFINITION	UNITS
6dof7	Six Degrees Of Freedom Model	
A	Area	ft² (m²)
D	Drag Force	lbf (N)
C_d	Coefficient of Drag	
CG	Center of Gravity For Aircraft	
F	Force	lbf (N)
L	Roll Moment about the aircraft X axis	lbf-ft (N-m)
M	Pitch Moment about the aircraft Y axis	lbf-ft (N-m)
N	Yaw Moment about the aircraft Z axis	lbf-ft (N-m)
V_T	Vector composing velocities $\begin{bmatrix} U & V & W \end{bmatrix}$	ft/s (m/s)
U	Velocity along the aircraft X axis	ft/s (m/s)
V	Velocity along the aircraft Y axis	ft/s (m/s)
W	Velocity along the aircraft Z axis	ft/s (m/s)
X	X axis position of coordinate system	ft (m)
Y	Y axis position of coordinate system	ft (m)
Z	Z axis position of coordinate system	ft (m)
b	Damping Constant	lbf-s/ft (N-s/m)
d	Diameter	ft (m)
k	Spring Constant	lbf/ft (N/m)
l	Length of strut	ft (m)
t	Thickness	ft (m)
u_∞	Free-stream Velocity Scalar, $\sqrt{U^2 + V^2 + W^2}$	ft/s (m/s)
q	Dynamic Pressure, $\frac{1}{2} * \rho * u_\infty^2$	lbf (N)
ρ	Density	slug/ft³ (kg/m³)
*	Any Legal Character String For Source Code File Name	

SUBSCRIPT	DEFINITION
atm	Earth Coordinates, Atmospheric
gear	Gear Coordinates
body	Body Coordinates
max	Maximum
strut	Gear Strut Value
tire	Gear Tire Value
T	True, Body Coordinates
t	Torsional Constant
x	X axis direction
y	Y axis direction
z	Z axis direction

3

INTRODUCTION

In an effort to decrease the linearity required for realistic simulation of an aircraft a six degrees of freedom model was adapted from the previous Simulink S-function code. This new addition would incorporate braking and handling while the aircraft was located on the ground. The simulator code also managed to include portions that allowed shock absorber simulation using spring and damping constants for the aircraft to land. Contrarily, one of the major shortcomings of this modeling was that it only tabulated forces while in contact with the earth and negated the effects of landing gear in flight.

In order to append to the work previously done by Chris Atkinson's senior project, the C++ code was further modified to allow force calculation both on the ground and while in flight. These additions would allow the results to not only add to the realistic quality of the Calpoly Flight Simulator but moreover demonstrate the effect of gear design and their consequences on aircraft behavior. Through assorted modifications of the gear arrays variables included in the text file, an infinite number of gear configurations are possible.

With inclusion of this module the next step was to incorporate a system that allowed the gear to trigger on and off depending on the Simulink model. While the previous version kept a constant gear down value, this technique of handing the landing gear would cause problems with a routine that integrated forces into the flight dynamics. In an effort to alleviate these issues, a trigger based on elevation was designed that would permit the landing gear to enable only when required. An addition to this model has also

4

accounted for the extension and retraction of the landing gear and the resultant on the forces.

Lastly, an Excel database capable of generating the necessary text files was created to reduce the time needed for rapid prototyping of various gear configurations. This would effectively reduce the required intricate knowledge required to modify the heavily structured files read by the C++ coding. The process has now been reduced to simply a matter of inputting the amount of gear needed to be modeled and the respective variables into an already formatted spreadsheet that generates when desired.

5

ANALYSIS

For proper initiation of the gear drag forces and their counterpart moments, the equations had to be generated along with their handling methods in order to code the model efficiently. First, the normally-used equation for generating drag across an aircraft was taken and manipulated to solve for the drag to be generated.

Coordinate Systems	
Gear Coordinate System	
X Axis	Forward in Direction of Rotation
Y Axis	Axis of Rotation
Z Axis	Along Strut
L Moment	Rotation about X Axis
M Moment	Rotation about Y Axis
N Moment	Torsional Rotation about Z Axis
Body Coordinate System	
X Axis	X Axis: Out Through Nose of Aircraft From CG
Y Axis	Y Axis: Out Along Right Wing From CG
Z Axis	Z Axis: Down From CG
L Moment	Roll, φ
M Moment	Pitch, θ
N Moment	Yaw, ψ
Area Coordinate System	
X Axis	Front Gear Area
Y Axis	Side Gear Area
Z Axis	Bottom Gear Area

Table 1: Coordinate Systems

While air density (ρ) was the only constant value for all axes required, true velocity (V_T) had to be imported from body coordinated to gear coordinates, and the necessary coefficients of drag (C_d) and area (A) numbers had to be read in.
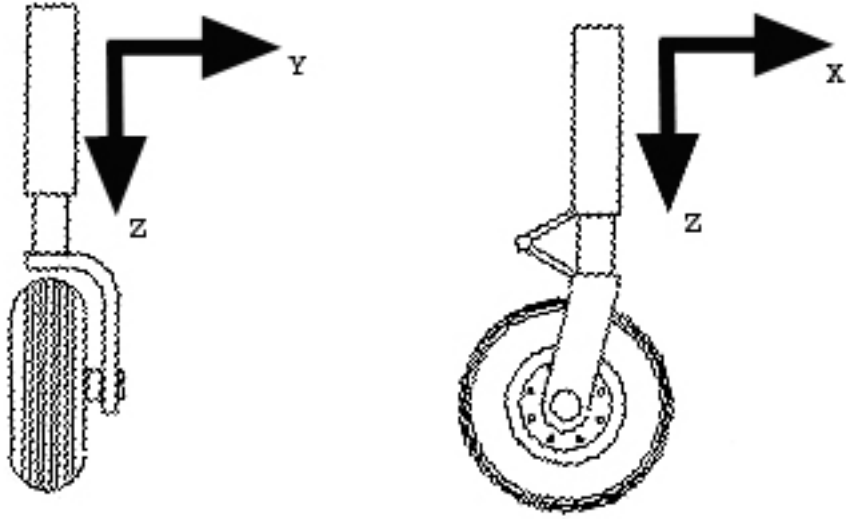
Figure 1: Gear Coordinate Axes

$$C_d = \frac{-D}{\frac{1}{2} * \rho * V_T^2 * A}$$

Equation 1: Drag Force Formula

Tackling the drag force required the breaking up of each area and coefficient of drag into their respective axes first and then using summation of all contributing drag creation areas into a main point force. Each summation is performed on the individual gear and then their contributions are related back to the main aircraft. While the equations for the drag on the Y and Z axis of the gear are similarly a combination of strut and tire drag, the Z axis strut drag is negated since only tire is exposed. Please note the drag forces are referenced according to the axis in which they act upon and not the area plane in which they act normal to (i.e. X drag force requires the frontal gear area projected onto the Y plane). The introduction of this unorthodox coordinate system has been created in order to resolve any issues concerning the representation of 3-D models in a 2-D environment and separate the area from planar designations.

7

$$D_x = -\frac{1}{2}*\rho*U^2*\sum(C_d*A)_x$$

$$D_y = -\frac{1}{2}*\rho*V^2*\sum(C_d*A)_y$$

$$D_z = -\frac{1}{2}*\rho*W^2*\sum(C_d*A)_z$$

Equation 2: Drag Force Summation on Singular Gear

$$D_x = -q_x*\left(\left(C_{d_x}*A_x\right)_{strut}+\left(C_{d_x}*A_x\right)_{tire}\right) \qquad q_x = \frac{1}{2}*\rho*U^2$$

$$D_y = -q_y*\left(\left(C_{d_y}*A_y\right)_{strut}+\left(C_{d_y}*A_y\right)_{tire}\right) \qquad q_y = \frac{1}{2}*\rho*V^2$$

$$D_z = -q_z*\left(C_{d_z}*A_z\right)_{tire} \qquad\qquad\qquad q_z = \frac{1}{2}*\rho*W^2$$

Equation 3: Drag Formulas for Gear Axes

$$A_{x\,strut} = A_{y\,strut} = d_{strut}*l_{strut}$$

$$A_{x\,tire} = A_{z\,tire} = t_{tire}*d_{tire}$$

$$A_{y\,tire} = \pi*\left(\frac{d_{tire}^2}{4}\right)$$

Equation 4: Gear Area Formulas

The formulas for the areas and their simplification to more elementary portions are depicted in Equations 4 and 5. Notice that the Y and X contribution for strut drag are completely identical. This result is due to the assumption that the strut is essentially cylindrical in nature along the Z axis. While a box-like configuration could be edited into the coding with very little difficulty, the result would overcomplicate the gear array with dimensions for all axes and their coefficients of drag. Rather then have the user identify six new variables, two now take their place in order to maintain usability for common users upon completion. The tire design itself is again simplified down to a smaller cylinder with rotational axis perpendicular to the ground with very little lost in this
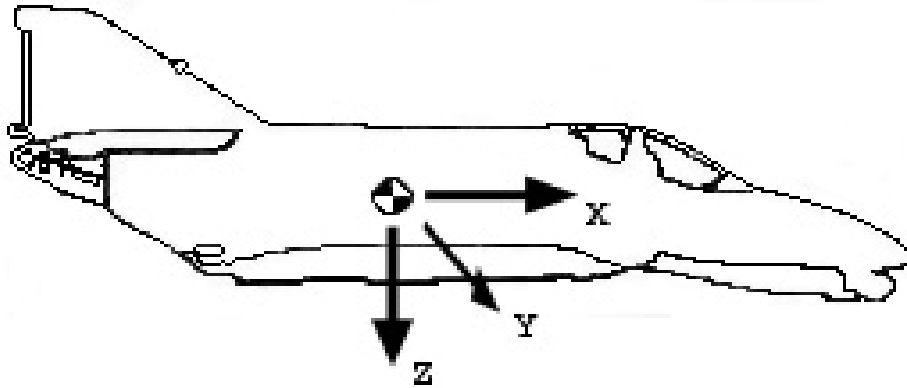
8

translation to code.

Figure 2: Aircraft Coordinate System

$$L = F_z * Y_{body}$$

$$M = F_x * Z_{body}$$

$$N = F_y * X_{body}$$

Equation 5: Force Relations to Aircraft Moments

Now with the forces rendered, the necessary couples on the aircraft are created by moving the point forces to the CG and multiplying them by their appropriate lever arms. The equations listed within Equations 6 relate the aircraft moments' effects to their causes on the landing gear conversion. For the conversion of the moments, special attention was paid to the lever arms since the moment would be affected drastically, i.e., reduced or increased through improper initiation. For calculations of the moment arm, first the coordinates are translated to the gear axis using the offsets defined by the user-specified values. Once translated to the zero axes of the gear coordinates, the centroid of area is calculated using the formulas in Equation 6 to determine the necessary

9

offset on the Z axis in which to apply the point drag force. The $\overline{Z}$ formula has been given a subscript to indicate the contributing area (X: frontal area or Y: side area).

$$\overline{X} = \frac{\sum_i A_i * \overline{x}_i}{\sum_i A_i} \qquad \overline{Y} = \frac{\sum_i A_i * \overline{y}_i}{\sum_i A_i}$$

Equation 6: Area Centroid Equation

No centroid equation is applied to the Z axis itself since it is an assumed rigid body with no bending displacement from gear alignment. Accounting for this bending effect in the X and Y axes would effectually overcomplicate the landing gear portion of the code beyond that of the processors capability. In addition, many of the spring and damping values for the shock absorber would need to be reevaluated at each time step due to the decreasing effectiveness of the shock absorber once contorted. The final equations denoted are the centroid calculations as measured from the zero points along the X, Y, and Z gear axis.

$$\overline{Z}_x = \frac{A_{x\,strut} * \dfrac{l}{2} + A_{x\,tire} * \left(l_{strut} + \dfrac{d_{tire}}{2}\right)}{A_{x\,strut} + A_{x\,tire}} \qquad and \qquad \overline{Z}_y = \frac{A_{y\,strut} * \dfrac{l}{2} + A_{y\,tire} * \left(l_{strut} + \dfrac{d_{tire}}{2}\right)}{A_{y\,strut} + A_{y\,tire}}$$

Equation 7: Area Centroid Calculation

$$\overline{Z}_x = \frac{[l_{strut} * d_{strut}] * \left[\dfrac{l}{2}\right] + [d_{tire} * t_{tire}] * \left[l_{strut} + \dfrac{d_{tire}}{2}\right]}{[l_{strut} * d_{strut}] + [d_{tire} * t_{tire}]}$$

Equation 8: X Area Centroid

$$Z_y = \frac{[l_{strut} * d_{strut}] * \left[\dfrac{l}{2}\right] + \left[\pi * \left(\dfrac{d_{tire}}{2}\right)^2\right] * \left[l_{strut} + \dfrac{d_{tire}}{2}\right]}{[l_{strut} * d_{strut}] + \left[\pi * \left(\dfrac{d_{tire}}{2}\right)^2\right]}$$

Equation 9: Y Area Centroid

10

PROCEDURE

Initiation of the changes necessary for completion of the task of adding gear

forces began by first editing the C++ S-function code that controlled the six degrees of

freedom model. Inclusion of the needed force generation, appropriate moments and new

variable declarations were completed within the new version. Afterward, a Simulink

model was then generated to allow the gear to toggle on and off as desired while allowing

a smooth transition between no forces and the maximum occurring by full extension of

the gear. Lastly, an Excel macro was formed that would handle the large array of

variables as well as the text file required for variable reading.

S-Function C++ Code

The capability of coding C++ to create entirely unique Simulink blocks expands

the capability beyond the confines of specified blocks. The process requires the two

necessary programs of Matlab 6.0 or higher and Visual C++ or another version of C++

compiler. Once the code is completed or a syntax error check is required, the source is

compiled in a method known to Matlab as "mexing." Mexing is the equivalent process of

compiling source code using a compiler built into Matlab that ensures inclusion of the

necessary headers a Simulink block requires. Headers, once included, allows the code to

include necessary commands for functionality as a Simulink block. Setting up for mexing

is launched by typing "*mex –setup*" into the Matlab command window. The setup allows

options to use an outside C++ compiler if available while still maintaining a link to the

proper headers. Once set up, the capability exists to convert any *.cpp file into a *.dll file,

the compiled C++ code that is the driving force behind S-functions. To accomplish the

11

actual compiling a user would type mex, a space, and then the given file name with file extensions *.c or *.cpp. At this point either a *.dll file will be produced or the Matlab prompt will list the various syntax errors that need correction. Once the *.dll file has been added to the working directory of the Pheagle Flight Simulator, a name change to the S-function block mask matching the file name is made to accept the new programming.

To begin, the source of the S-function "sixdof7.cpp" was first opened and then examined closely for an optimal position to insert the new routine. Upon analysis it was determined that the desired forces must be independent of whether the aircraft was airborne or touching earth. For this reason it was opted to insert the module just before the Boolean based if-then statements controlling the gear up and down handling. From this point, transitions to the gear- and tire-based coordinated systems are made and drag force calculations are performed using standard drag formulas.

The velocities are handled in such a manner as to allow breakup of the aircraft velocities onto the three major gear axes. This approach seemed more practical than a simplification to a singular scalar velocity acting upon the gear or consideration of only frontal effects. After computations for forces are accomplished, the forces are transitioned to the origins based on gear geometry with the proper moment creation accounted for. Since a process contained at the end of the ground contact code accounts for translation of these formed dynamics from gear coordinated to the main aircraft body, the procedure was moved outside the if-then operation to minimize effective changes. To

12

conclude the modifications, the newly needed variables were added to the gear text file reading parameters contained in the model initialization coding.

Simulink Ratio Model

To begin the process, an altitude trigger was installed into the model that could enable or disable the gear based on height above sea level. Once the function of gear extension and retraction outputs had been installed, a second layer subsystem was placed to handle the gradual transfer of the gear. This was deemed necessary due to force feedback from the Pheagle stick and the shock that would occur when the gear drag forces occurred.
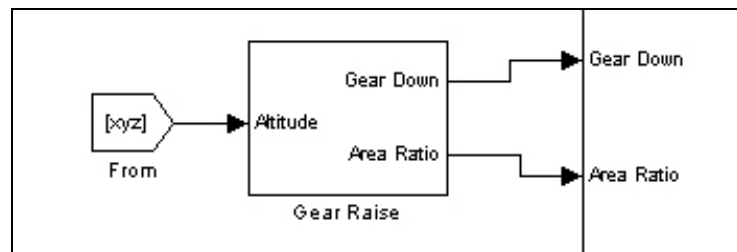

Figure 3: Mask Layer of Gear Trigger Model

Modifying the Simulink model further to accommodate an area ratio input required an evaluation of the inputs that would be sent to the six degrees of freedom model. For an initial start-up of the model, the gear output was set to one. During flight the gear output program then shifts, due to the altitude trigger, between one for gear fully extended and zero for gear fully retracted. To accomplish this task transfer functions, if subsystems, constant blocks, merge blocks, and relational operators were used in an effort to replicate such a signal. A transfer function was decided on mainly for the purpose of slowing down the signal between full gear extension and retraction to the gear

13

forces. For this reason, the first order actuator transfer function $\dfrac{0.5}{s+0.5}$ was selected to

opt as the area ratio signal since it created a believable settling time and rise time for a

motor driven system.

One issue arose over model initialization since the model required time to slowly

adjust to a solid signal of one for model start-up. In an effort to correct this oversight, an

if-action subset was created to allow the transfer function to slowly correct from the

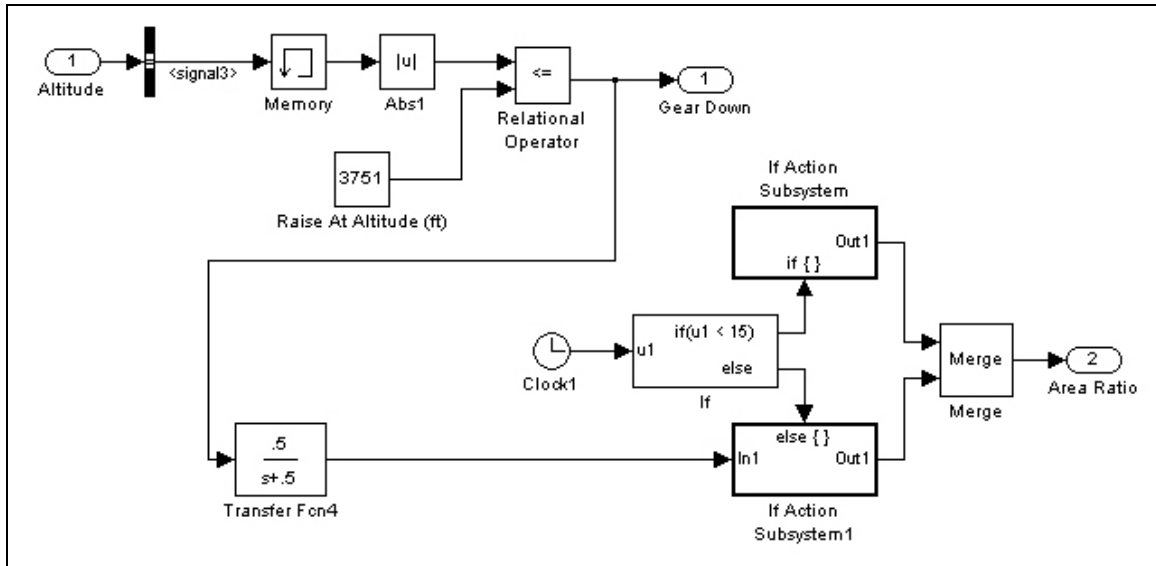initial output of zero from the flight simulator.

Figure 4: Area Ratio Coding and Initialization Parameters

Excel Macros Table Generator

Excel by nature is a powerful engine for calculations and formatting because of

the cell structuring of databases. When coupled with the programming ability of Visual

Basic, it provides the ability to perform tasks outside the range of mere data tracking. In

an effort to enable quick construction of the various text files required by the flight

14

simulator, an effort was put forth to create an method that would output specified data into a tab delimited document. To access the macro ability of Excel, a Microsoft Visual Basic or similar program should be installed to make full use of modules and various other functions. Once completed, a right click on any worksheet and the selection of view code is adequate to bring the user into the Visual Basic editor.

To aid in the process the Visual Basic design toolbar can be used to add event driven buttons and toggles. Two buttons were used in the design: one to layout necessary formatting and the second to create the text file upon the user's desire. Upon completion of the landing gear output generation code, a master worksheet was created for the addition of other buttons that handled some of the other important flight simulator text files. Toggling between the master sheet and new gear forms allows the user to work independently of established data. This avoids the possibility of deletion or modification of required data for model operation and experimentation that can be performed by the user.

With gear configuration complete, other features for the flight simulator were added that allowed the text generator to also create the broadcast IP address, gear arrays, and crash coordinate texts. Another inclusion was a module to handle writing to the other flight simulator computers, allowing the possibility to write necessary files remotely. While the designs were not essential, they proved extremely helpful and the progress was substantial enough to note as proof of working operation.

DISCUSSION OF RESULTS

The evaluation of the modified gear code has produced interesting, although expected, results. The drag forces introduced onto the gear provided a distinct negative pitching moment while extended. Testing has proven that the aircraft, when trimmed into straight and level flight below the altitude gear-up marker, will cause a positive pitching moment once the gear is retracted. As expected, by means of the essential drag equations, the forces increased exponentially as velocity increased. Interestingly enough, this event occurred a significant amount only on the gear's X axis, while its effect on the Y and Z axes could be considered negligible in comparison. This is most likely attributed to the extreme yaw and roll required to get any noteworthy changes along the Y axis, while the Z axis would require a stall of equal proportions.

Concerning the file read values used, it was decided to use raw area and coefficient of drag values rather than approximated aircraft geometry calculations. While the mathematics involved for assigning a specific configuration for the strut and tire areas would require no additional text file changes, loss in the configurability of the simulator would be sacrificed. Instead, it was opted to maintain a user input based value for each gear, allowing easier adaptation to various experimental gear constructions. Such gear constructions would include different gear types to be enabled for each leg to test handling quality. Results from such experimental drag relations for each leg have confirmed that distance from the center of gravity of the aircraft has proven the most dominant force. Providing a larger lever arm for the gear forces to operate created a massive rotation for the aircraft to correct from and required greater pilot adjustment.

16

The addition of the additional input parameter for area ratio proved important in the final design, since testing these new forces would enable feedback on the Pheagle stick. The area ratio model proved a valuable asset in reducing the shock and load placed onto the electromagnets controlling the force feedback. This module aided in determining values for coefficients of drag and areas that produced effects within reasonable margins. Also, the coding of this parameter allowed a basis on another factor aside from Boolean input. The gradual transition helped prevent sudden shifts in the aircraft's dynamics thereby eliminating the need for massive pilot adjustments or unreasonable moments. The trigger driving the introduced area ratio, while still based on altitude calculation within the model, is a configurable parameter capable of managing gear extension or retraction triggering with exceptional ease. In addition, such a trigger can be recreated to allow a hardware switch mounted in the cockpit to act as the control with very little modification.

While the source of more realistic results, the area ratio model has also proved to be the most problematic task of the entire operation. By nature of a transfer function, and the reason for its use, requires time based on the denominator and numerator setup to arrive at the desired value. Unfortunately, this conflicts with the flight simulator's ability to reset and initialize properly since the initial condition is 0 for these abilities. This becomes problematical because the flight simulator now needs time to set the value to 1. Only a quick fix was managed through the use of if-then blocks. A more in-depth, permanent solution would require yet another C++ based transfer function capable of including the reset triggering. The task was deemed unreasonable during the stages of

17

development of the gear forces since operation of the forces and moments was considered

the primary task, but this oversight will be resolved at a later time.

CONCLUSION

The addition of gear drag forces, and modifications to the Simulink model, while very minute, are essential for the creation of a realistic flight simulator. Various changes were noted in the flying capability of the Calpoly Pheagle Flight Simulator, including drastic changes in pitching during gear extension. With these extensions to the gear module, the true depth of the aircraft is revealed and the knowledge gained can now allow further additions with ease. The addition of the Excel Table Generator has already proven its worth between configuration switches and Simulator setup. While the functionality is still limited, with future work the engine could be capable of creating all the necessary values in addition to the formats, rendering it an extremely valuable calculation design tool.

Although the best effort was made to prevent irregularities from being adopted into the code, some unrealistic results have shown themselves in assorted places. Many of these problems find their origin in the use of a transfer function to enable a nonlinear control over the landing gear. The findings support that this may be an issue even with external hardware control of the mechanism. Only a properly instituted switch with another C++ coded S-function will truly solve the difficulty.

19

APPENDIX A: Gear Configuration Text

With alterations to the coding of the six degrees of freedom model controlling the flight simulator came changes in the gear configuring array. The most notable of these additions are placements for strut and tire areas as well as drag coefficient values.

Landing Gear Variables Listing		
1	X_{body}	X Location Forward From Aircraft CG (ft)
2	Y_{body}	y Location Right From Aircraft CG (ft)
3	Z_{body}	Z Location Down From Aircraft CG (ft)
4	k_x	X Gear Spring Constant (lb/ft)
5	k_y	Y Gear Spring Constant (lb/ft)
6	k_z	Z Gear Spring Constant (lb/ft)
7	b_x	X Gear Damping Constant (lb-s/ft)
8	b_y	Y Gear Damping Constant (lb-s/ft)
9	$b_{z\,i}$	Z Initial Gear Damping Constant (lb-s/ft)
10	$b_{z\,f}$	Z Finial Gear Damping Constant(lb-s/ft)
11	X_{tire}	Wheel Offset Forward (ft)
12	k_t	Gear Torsional Spring Constant(ft-lb/rad)
13	b_t	Gear Torsional Damping Constant (ft-lb-s/rad)
14	$I_{z\,t}$	Z Mass Moment Of Inertia (slug-ft²)
15	$absorb$	Use Strut (1 = yes, 0 = no)
16	z_{strut}	Stroke Length (ft)
17	D_{strut}	Piston Diameter (ft)
18	P_{ext}	External Pressure (lb/ft²)
19	P_{int}	Internal Pressure (lb/ft²)
20	γ	Polytrophic Compression Exponent
21	k_{tire}	Tire Spring Constant (lb/ft)
22	$k_{t\,tire}$	Tire Torsional Spring Constant(ft-lb/rad)
23	$A_{contact}$	Tire Contact Area (ft²)
24	$F_{x\,max}$	Maximum X Gear Force (lb)
25	$F_{y\,max}$	Maximum Y Gear Force (lb)
26	$F_{z\,max}$	Maximum Z Gear Force (lb)
27	d_{tire}	Tire Diameter (ft)
28	t_{tire}	Tire Thickness (ft)
29	$C_{d\,strut}$	Strut Coefficient Of Drag (ft^2)
30	$C_{d\,x}$	Front X Tire Coefficient Of Drag
31	$C_{d\,y}$	Side Y Tire Coefficient Of Drag
32	$C_{d\,z}$	Bottom Z Tire Coefficient Of Drag

Table 2: Gear Variable Array

For each piece of landing gear the following 32 variables are necessary for operation and none of which can be omitted. These values classify as orientation,

20

dimensioning, dynamics, or drag characteristics for the individual pieces and can be

unique to each leg. A reminder, when translating this to actual code the C++ counting

system for all text files associated with the "6dof7" model begin at 0 thus the output from

the table creator will show only 31 as its final number.

APPENDIX B: Table Creator Source Code

The written text exporter code follows a simple format and is divided into two critical parts. The first section sets up the needed variables to write the specified block to a text file, while the second portion performs the routine of writing the given section to a text file of specified name. The subdivision of content and the writing function allows further additions of text generation buttons so long as the proper variables are specified.

```
Private Sub CommandButton1_Click()

    ' Create F-4 Gear Text
    ' Variables are Starting Column, Starting Row, Ending Column, Ending Row, Sheet
name, Text name

    TableCreation 1, 3, 5, 89, "F-4", "SimGear.txt"
    Worksheets("Master").Range("I26") = "F-4 Phantom Gear File"

End Sub
```

Figure 5: Call Command for Text Write Function

The writing routine itself relies on for-statements to read through each row to a specified column, print the contents to a file, and then move to the next row. The process repeats until the last row is reached, whereupon a blank line is written and the file is closed. The path and file name must be specified prior to writing the file but, since each specified range of cells corresponds to one file in a specific location in the Flight Simulator file structure, the predetermined path allows files to be generated only where needed. In addition, with error checking removed the files are allowed to overwrite existing files of the same name thus not requiring reconfiguring of the actual Simulink blocks. As long as the data for the Pheagle is stored into a cell array, it too can be called and written using the same button call as specified above.

22

```vb
Sub TableCreation(ColStart As Integer, RowStart As Integer, ColEnd As Integer, RowEnd As
Integer, Shtname As String, Destfile As String)
' File creation routine for tables
' Dimension all variables
Dim FileNum As Integer
Dim ColumnCount As Integer
Dim RowCount As Integer
Dim Path As String

    ' Specified path in which to write to
    Path = "\\Crewchief\c\Pheagle 4.0\Workingdir\Latest Versions\"

    ' Obtain next free file handle number
    FileNum = FreeFile()
    ' Turn error checking off
    On Error Resume Next

    ' Attempt to open destination file for output
    Open Path & Destfile For Output As #FileNum

        ' If an error occurs report it and end
        If Err <> 0 Then
        MsgBox "Unable To Create File: " & Filename, vbCritical, "File Status"
        End
        End If

    ' Turn error checking on
    On Error GoTo 0

    Worksheets(Shtname).Select
    Worksheets(Shtname).Range("A1").Select

        ' Loop for each row in selection
        For RowCount = RowStart To RowEnd
            ' Loop for each column in selection
            For ColumnCount = ColStart To ColEnd

            ' Write current cell's text to file
            Print #FileNum, Selection.Cells(RowCount, ColumnCount).Text;

                ' Check if cell is in last column
                If ColumnCount = ColEnd Then
                ' If so then write a blank line
                        Print #FileNum,
                Else
                ' Otherwise tab over
                Print #FileNum, Tab;
                End If

            ' Start next iteration of ColumnCount loop
            Next ColumnCount
        ' Start next iteration of RowCount loop
        Next RowCount

    'Close destination file
    Close #FileNum

    ' Return Back to master switchboard and return success status
    Worksheets("Master").Select
    MsgBox "File Has Been Successfully Created: " & Destfile, , "File Status"
End Sub
```

Figure 6: Text File Write Routine

APPENDIX C: Table Creator Sample Output File

The file generated from the table creator corresponds precisely with the format required from the S-function read in. To allow for future reference and configuring of the gear text file the values currently used are displayed below within the table structure of Figure 7. Note that the cell values must be tab delimited when copied from this table to maintain program readability.

3		Number of Pieces of Landing Gear
(0) Back Right Leg ------------------		
-2.9	0	x Location forward (ft)
9	1	y Location right (ft)
8.5	2	z Location down (ft)
20000	3	kx (lb/ft)
20000	4	ky (lb/ft)
0	5	kz (lb/ft)
4000	6	bx (lb s/ft)
4000	7	by (lb s/ft)
3000	8	bz Initial (lb s/ft)
8000	9	bz Final (lb s/ft)
0	10	Wheel offset forward (ft)
5000	11	kt (ft lb/rad)
2000	12	bt (ft lb s/rad)
100	13	Iz (slug ft^2)
1	14	Use Strut (1 - yes, 0 - no)
1.32	15	Stroke (ft)
0.25	16	Piston Diameter (ft)
2116.8	17	External Pressure (lb/ft^2)
55237	18	Internal Pressure (lb/ft^2)
1.2	19	Polytropic Compression Exponent
120000	20	Tire Spring Constant (lb/ft)
5000	21	Tire Torsional Spring (ft lb/rad)
0.35	22	Tire Contact Area (ft^2)
120000	23	Maximum x Force (lb)
150000	24	Maximum Y Force (lb)
190000	25	Maximum z Force (lb)
3	26	Diameter of Tire (ft)
1	27	Tire thickness (ft)
0.1	28	Strut coefficient of Drag
1.18	29	Tire Frontal coefficient of Drag
1.16	30	Tire Side coefficient of Drag

24

1.18	31	Tire Bottom coefficient of Drag

(1) Back Left Leg ------------------

-2.9	0	x Location forward (ft)
-9	1	y Location right (ft)
8.5	2	z Location down (ft)
20000	3	kx (lb/ft)
20000	4	ky (lb/ft)
0	5	kz (lb/ft)
4000	6	bx (lb s/ft)
4000	7	by (lb s/ft)
3000	8	bz Initial (lb s/ft)
8000	9	bz Final (lb s/ft)
0	10	Wheel offset forward (ft)
5000	11	kt (ft lb/rad)
2000	12	bt (ft lb s/rad)
100	13	Iz (slug ft^2)
1	14	Use Strut (1 - yes, 0 - no)
1.32	15	Stroke (ft)
0.25	16	Piston Diameter (ft)
2116.8	17	External Pressure (lb/ft^2)
55237	18	Internal Pressure (lb/ft^2)
1.2	19	Polytropic Compression Exponent
120000	20	Tire Spring Constant (lb/ft)
5000	21	Tire Torsional Spring (ft lb/rad)
0.35	22	Tire Contact Area (ft^2)
120000	23	Maximum x Force (lb)
150000	24	Maximum Y Force (lb)
190000	25	Maximum z Force (lb)
3	26	Diameter of Tire (ft)
1	27	Tire thickness (ft)
0.1	28	Strut coefficient of Drag
1.18	29	Tire Frontal coefficient of Drag
1.16	30	Tire Side coefficient of Drag
1.18	31	Tire Bottom coefficient of Drag

(2) Nose Leg ----------------------

21.2	0	x Location forward (ft)
0	1	y Location right (ft)
8.5	2	z Location down (ft)
20000	3	kx (lb/ft)
20000	4	ky (lb/ft)
0	5	kz (lb/ft)
4000	6	bx (lb s/ft)
4000	7	by (lb s/ft)
800	8	bz Initial (lb s/ft)
2000	9	bz Final (lb s/ft)
-0.5	10	Wheel offset forward (ft)

25

1000	11	kt (ft lb/rad)
1000	12	bt (ft lb s/rad)
50	13	Iz (slug ft^2)
1	14	Use Strut (1 - yes, 0 - no)
1.32	15	Stroke (ft)
0.132	16	Piston Diameter (ft)
2116.8	17	External Pressure (lb/ft^2)
55237	18	Internal Pressure (lb/ft^2)
1.2	19	Polytropic Compression Exponent
120000	20	Tire Spring Constant (lb/ft)
5000	21	Tire Torsional Spring (ft lb/rad)
1	22	Tire Contact Area (ft^2)
110000	23	Maximum x Force (lb)
110000	24	Maximum Y Force (lb)
140000	25	Maximum z Force (lb)
3	26	Diameter of Tire (ft)
1	27	Tire thickness (ft)
0.1	28	Strut coefficient of Drag
1.18	29	Tire Frontal coefficient of Drag
1.16	30	Tire Side coefficient of Drag
1.18	31	Tire Bottom coefficient of Drag

Table 3: Sample Write Routine Output

The current configuration listed relies on three pieces of gear: the nose leg, the back right leg, and back left leg. The can support the addition of more gear but limitations on the Simulink model itself requires modifications to include these additional gear. For this reason only the setup for these mandatory pieces are included but with minimal modifications the code can accept the current model maximum of ten. Arrangements in excess of ten landing gear legs become too bulky for the model to support and thus cause processing delays that affect the real time simulation. In addition, to maintain rapid simulation objectives three is also coded into the table system in order to minimize Simulink model modifications necessary. Alterations from this standard three landing gear pattern result in mandatory reassigning of the input arrays for the braking, steering, and wheel angle as listed in the Simulink model input table.

26

APPENDIX D: Simulink Model Parameters

Port Number	Input Data Title	Units	Range
1	Aircraft Weight	lbm	1
2	Controls		$n_{controls}$
3	Air Density, ρ	slugs/ft^3	1
4	External Body Forces, F_x, F_y, F_z	lbf	3
5	External Body Moments, L, M, N	ft-lbf	3
6	Atmospheric Earth Velocity, $V_{x\ atm}$, $V_{y\ atm}$, $V_{z\ atm}$	ft/s	3
7	Atmospheric Earth Acceleration, $\dot{V}_{x\ atm}$, $\dot{V}_{y\ atm}$, $\dot{V}_{z\ atm}$	ft/s	3
8	Initial Earth Position, X, Y, Z	ft	3
9	Initial Euler Angles, Ψ, θ, φ	rad	3
10	Initial Body Velocity, u, v, w	ft/s	3
11	Initial Body Rates, p, q, r	rad/s	3
12	Speed of Sound, a	ft/s	1
13	Coefficients of Friction, μ_s, μ_k		2
14	Brake Force	lbf	n_{gear}
15	Wheel Angle	rad	n_{gear}
16	Steering Engaged (1 = yes, 0 = no)	Boolean	n_{gear}
17	Gear Down (1 = yes, 0 = no)	Boolean	1
18	Area Ratio		1
19	Reset (1 = yes, 0 = no)	Boolean	1

Table 4: Simulink Input Ports

In Tables 4 and 5, "n" refers to a user defined number based on the specifications set forth in the appropriate text file. The gear parameters are contained within the gear configuration text file while controls and parameters are provided in the table of coefficients. For the remainder of the range values, the numbers designate the width of the signal required for operation of the model. Neglecting these critical port widths with result in an error within the model and a prompt requiring alteration until the matter is corrected. If issues arise over the operation of the model based on a "6dof7" block error, check to ensure all port widths are assigned properly based on the desired configuration.

27

Port Number	Output Data Title	Units	Range
1	Earth Position, X, Y, Z	ft	3
2	Euler Angles, Ψ, θ, φ	rad	3
3	Body Velocity, U, V, W	ft/s	3
4	Body Airspeed Velocity, U_T, V_T, W_T	ft/s	3
5	Body Rates, p, q, r	rad/s	3
6	Angle-of-Attack & Sideslip Angle, α, β	rad	2
7	Load Factor	g's	1
8	Force & Moment Derivatives, C_L, C_D, C_y, C_l, C_m, C_n		6
9	User-Defined		n_{param}
10	Landing Gear Body Forces, $F_{x\ gear}$, $F_{y\ gear}$, $F_{z\ gear}$	lbf	3
11	Landing Gear Body Moments, L_{gear}, M_{gear}, N_{gear}	ft-lbf	3
12	Maximum X Gear Force	lbf	n_{gear}
13	Maximum Y Gear Force	lbf	n_{gear}
14	Maximum Z Gear Force	lbf	n_{gear}
15	Wheel Angle	rad	n_{gear}
16	On Ground (1 = yes, 0 = no)	Boolean	1
17	Gear Fail (1 = yes, 0 = no)	Boolean	1
18	Crash (1 = yes, 0 = no)	Boolean	1

Table 5: Simulink Output Ports

For the units, the majority of the values designated as English system measurements with the exception of those marked as "Boolean". Boolean is a term denoting a value of 1 or 0 and fundamentally acts as a switch with an on or off position. While the program is the capable of allowing the ports to accept values other then the designated 1 or 0, the values are converted internally to Boolean using the reference of greater then or less then 0.5. The use of Boolean is very common and appears predominately in buttons or triggers whether it be hardware or software. Caution must be exerted when using values of this nature while programming in Simulink since the on position must be constant to create a lasting effect unless programmed otherwise. In

28

addition, these signals create sudden changes that must be accounted for within the Simulink model. The limitation of the signal to merely on or off can create sudden shocks to the model and can cause damage to flight simulator components if used carelessly. To avoid these issues, ensure that a gradual signal is produced for variables controlling flight performance or hardware using either a coded S-function, or the transfer functions and rate limiter blocks provided by Simulink.

Heat Collection Element Glass to Metal Seal Redesign
For Solar Power Generation Plant

Andrew Davol
John Allen

Mechanical Engineering Department
Cal Poly State University

Heat Collection Element Glass to Metal Seal Redesign
For Solar Power Generation Plant

Dr. Andrew Davol, PE[1]
John Allen[2]

Abstract

KJC Operating Company operates three solar energy plants in the Mojave Desert that couple large solar arrays to a conventional steam plant for power generation. KJC has experienced repeated failures of one very expensive element in their system. This element is a glass tube used to shield a stainless steel pipe, which carries a heat transfer fluid through the collectors. This glass tube is subjected to extremely high temperatures and must hold a vacuum to limit the heat transfer from the system. The joint at the end of the glass tube has specifically been a problem. Failure of these elements has cost KJC upwards of $2,000,000 per year. The scope of this work as outlined in the original proposal was to develop new design concepts, perform detailed analysis on all feasible concepts concentrating on the coupled thermal structural interaction at the joint that is failing, and test critical elements of the proposed designs including high temperature sealing techniques. The focus of this project changed significantly early on as it became clear in our initial investigations that no one had a clear idea of exactly what temperatures the problem joint was being subjected to. As a result this phase of the project mainly focused on developing heat transfer models that could accurately predict the temperature distribution in the joint under the varying conditions seen throughout the year. Test data was gathered on site with the support of KJC operating company to validate the models. Predictions of worst-case temperature distributions were then generated using the models.

Preliminary analytical work has been completed for a few design alternatives based on the computed temperature distributions. This work looks at possible solutions in a qualitative fashion to determine if further exploration is warranted.

Background

KJC Operating Company provides maintenance, repair, and operation services for the five Solar Electric Generating Systems (SEGS) located in the Mojave Desert near Kramer Junction, CA. Together the five SEGS plants produce 150 MW of power, enough to supply energy to over 50,000 homes using the sun as the primary energy source.

Each of the SEGS has a large solar field composed of rows of parabolic trough solar collector assemblies and a conventional power plant. The collectors are equipped with reflector panels, which track the movement of the sun using sun sensors and microprocessors. The sunlight is focused onto a heat collection element (HCE), a

---

[1] Assistant Professor, Department of Mechanical Engineering, California Polytechnic State University

[2] Graduate Student, Department of Mechanical Engineering, California Polytechnic State University

1

specially coated steel tube surrounded by vacuum-insulated glass, through which a heat transfer fluid flows. This fluid is heated in the solar field to temperatures of 390 ^{0}C and pumped through a series of heat exchangers to produce the superheated steam, which powers the turbine generator. The SEGS are hybrid plants and can utilize natural gas to supplement steam generation on cloudy days. A schematic of a plant is shown in Figure 1.
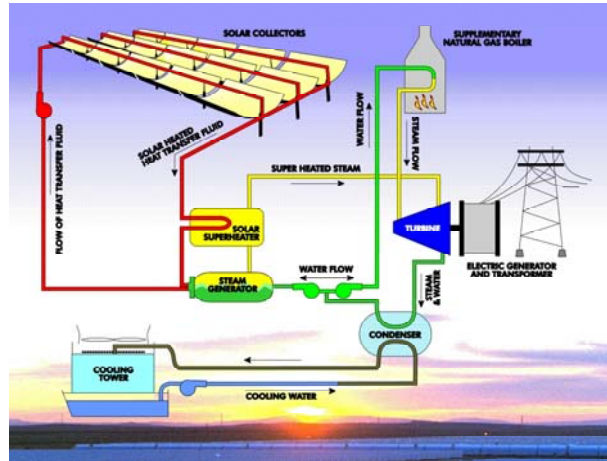

Figure 1, Schematic of KJC Solar Power Operation

The work presented in this document is focused on the design of the HCE glass to metal seal. Figure 2 below depicts the details of the HCE design in current use. The main components are the internal stainless steel pipe which is specially coated to aid in absorbing solar energy, the outer glass envelope used to hold a vacuum around the steel pipe to limit heat loss through convection, and a glass to metal seal used to attach the glass to the steel pipe through an expansion bellow that accounts for the thermal expansion difference between the glass and the steel.
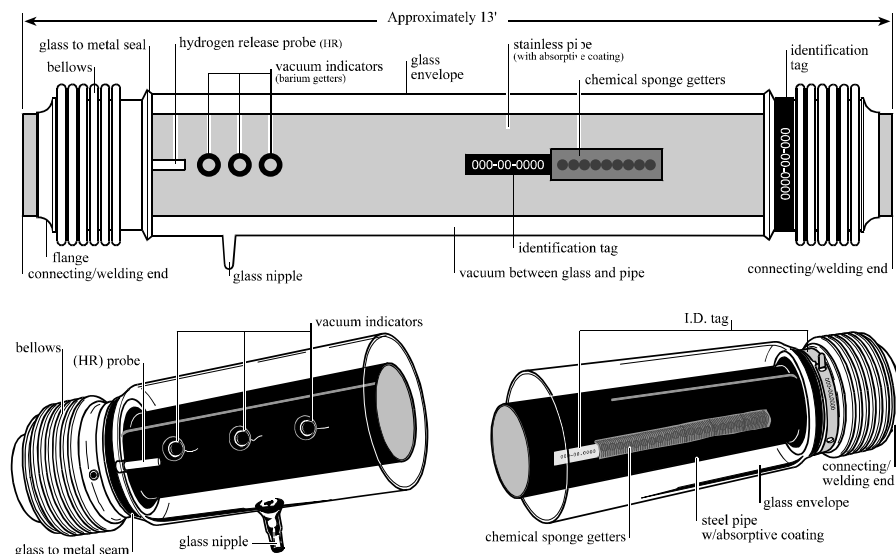
Figure 2, Current HCE Connection

2

Failure of the glass tube at the glass to metal seal has been a significant problem for KJC. As many as 2000 tubes a year have been replaced at a cost of approximately $1,000 apiece. Of the total failures 80% occur at the glass to metal joint (Figure 3). This is obviously a significant cost to the plant operator. Damage can also occur to the collector panels from falling glass. It has been postulated by KJC that the difference in the thermal expansion coefficient of the glass and the stainless steel connector is causing stresses in the glass that combine with the structural stresses to cause failure of the element.
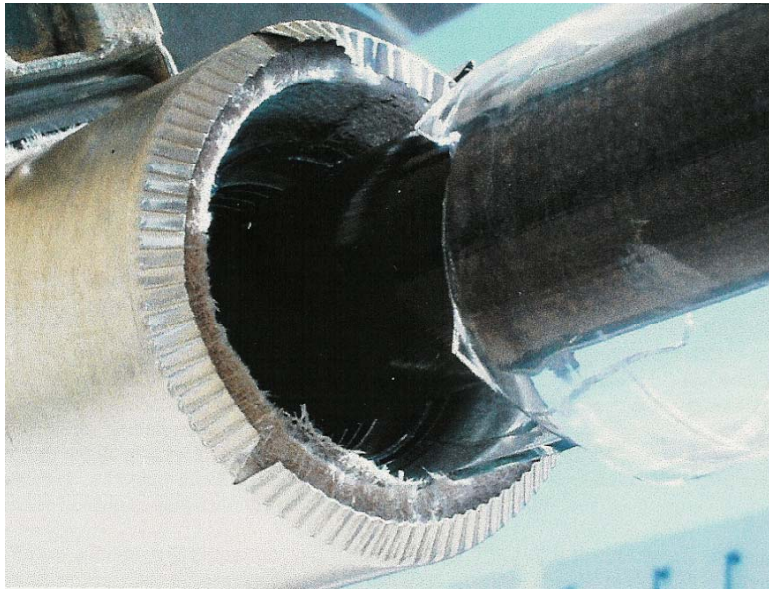
Figure 3, Failure of the HCE at the Glass to Metal Joint

The breakage of the HCEs is not constant throughout the year. Figure 4 shows the breakages per month for 2002. Spring seems to be the high season for breakage. It has been postulated that the lower sun angle in the spring is causing problems for the joint. This phenomenon is depicted in Figure 5. KJC has implemented an internal radiation shield in the newer HCE design that eliminates this problem. The new design in shown in Figure 6. While this fix seems to help, breakages are still occurring with the new design. Heat transfer models developed later in this document are used to predict the effect of differing sun angles on the joint temperatures.

Three design alternatives were presented in the original proposal. Of these three, two were mechanical connections and the third was a modification of the existing joint utilizing a composite material with an extremely low coefficient of thermal expansion to reduce the thermally induced stresses in the joint. It was decided that this stage of the project would focus on the third alternative, as it was the most feasible design to implement, and also look at alternate metal designs.

3
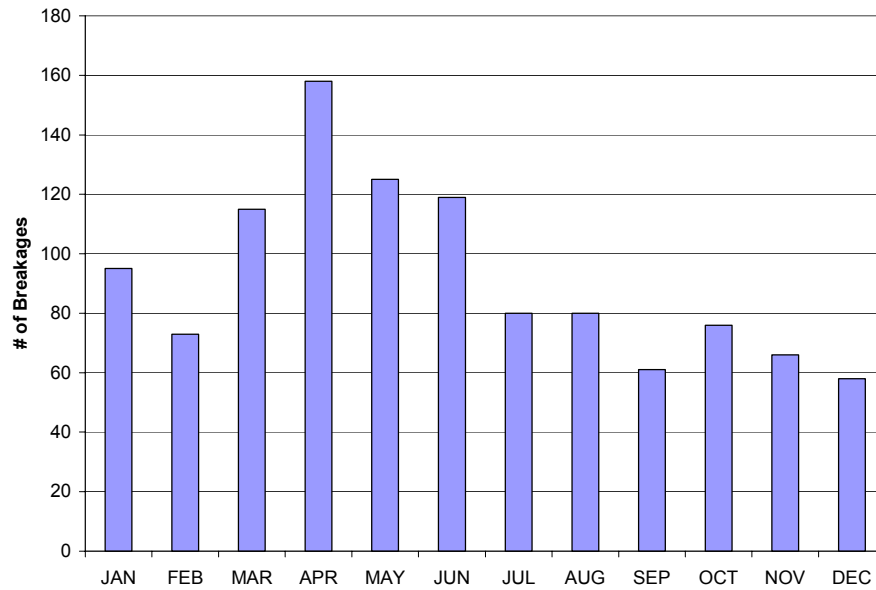
Figure 4, Breakage of HCEs for 2002
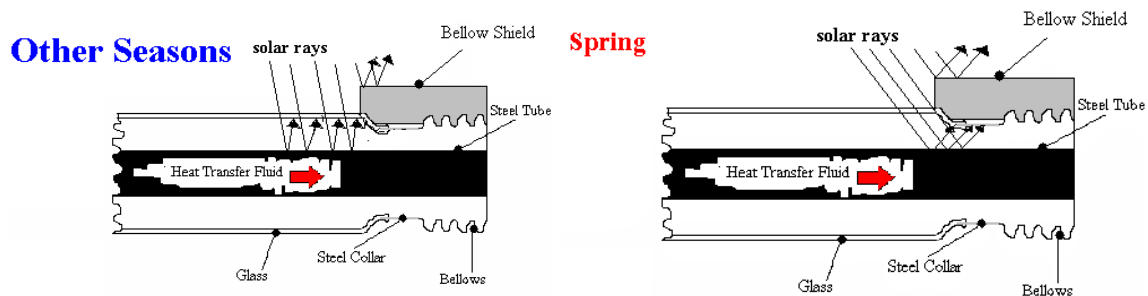
Figure 5, Sun Angle Effect on Joint Temperature
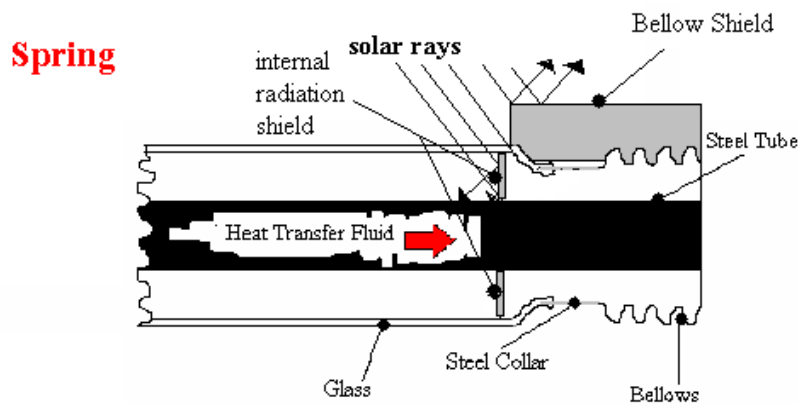
Figure 6, HCE with Radiation Shield

4

Testing

Testing was performed onsite at KJC to determine the temperatures in the joint region of the HCE during normal operation. Thermocouples were used to monitor the temperatures at various locations along the HCE and at several different positions circumferentially. The test positions are shown in Figure 7. Due to limitations in the data logger only 8 positions could be monitored at a time. Two separate 24 hour periods were monitored 26 days apart. Table 1lists the thermocouple positions monitored for each test.
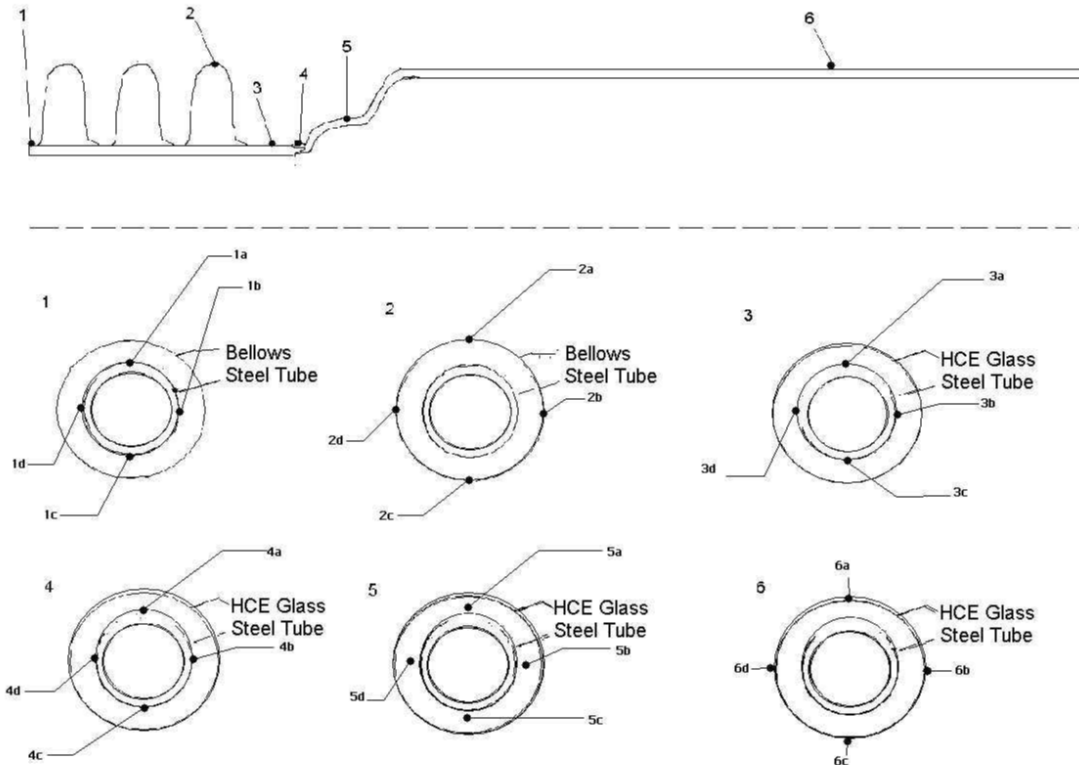
Figure 7, Thermocouple Positions

Table 1, Thermocouples Monitored

Date	Thermocouples Monitored
August 3rd	3a,3c,4a,4b,4c,4d,5a,5c
August 29th	1a,2a,3a,4d,5a,6a,6c,6d

It can be seen that there is significant temperature variation throughout the day. Figure 8 shows the temperature profile for position 5a for the 24 hour period monitored on August 29th. Modeling presented later in this report will focus on attempting to match the "steady state" temperatures at approximately 3:00pm (15:00). The results from this testing showed some variation in the temperatures circumferentially around the joint under the insulation. The change from the focal point side of the HCE to the far side around the seal

5

was approximately 15-20 °C. The longitudinal temperature distribution along the HCE is presented in Figure 9 for various times of the day. Figure 10 shows the variation in surface temperature for the instruments at position 6 from the August 29th test. The scatter in this data is mainly due to the variation in the wind speed throughout the test.

The following sections will present analytical models that attempt to match the temperature profile measured in these tests.
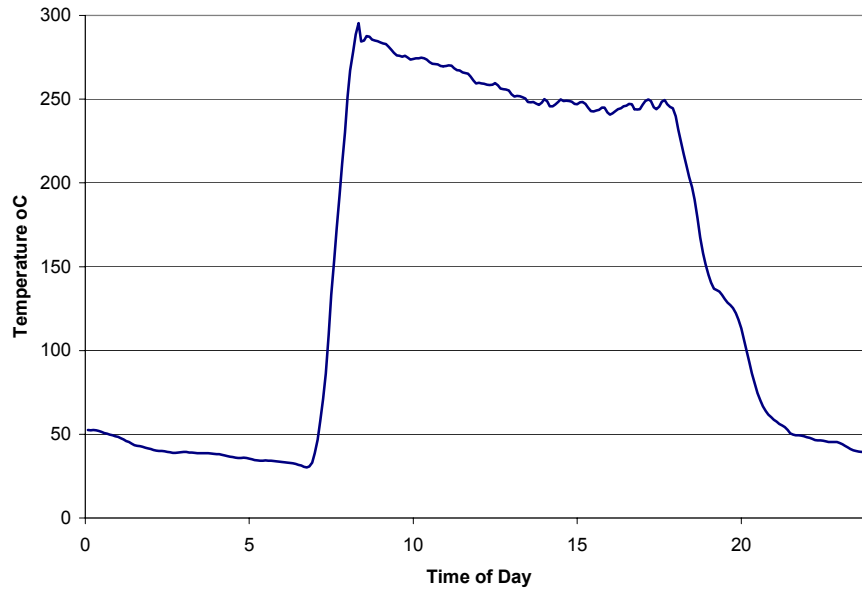
Figure 8, Temperature Profile at Position 5a an August 29th
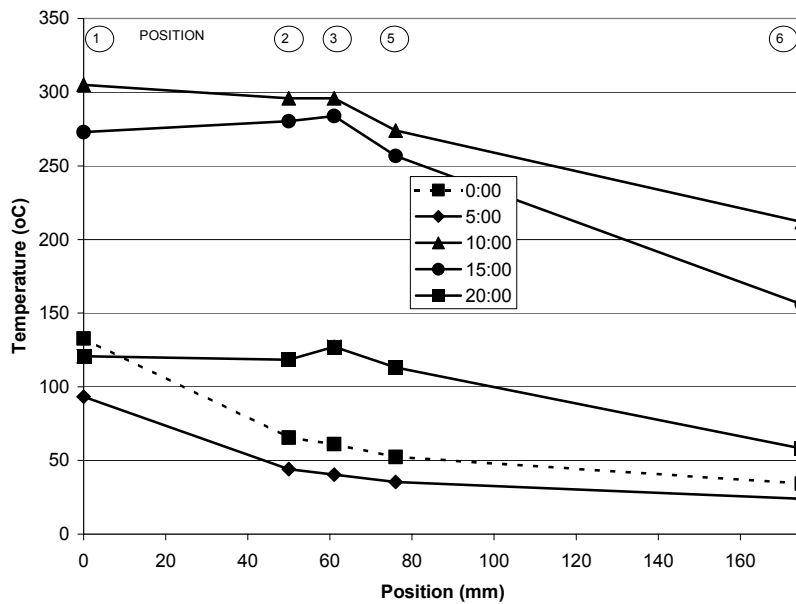
Figure 9, Longitudinal Temperature Distribution

6
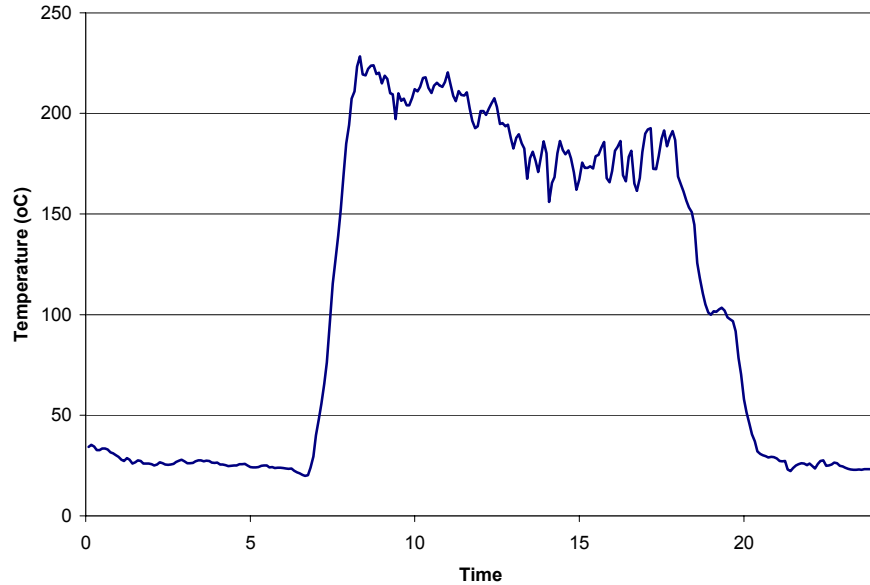
Figure 10, Glass Temperature at Position 6

Joint Temperature Distribution, Heat Transfer Models

Initial investigation into this problem showed that the maximum temperature originally quoted for this joint of 1090 C^o was well above the softening point of the glass. The focus for this stage of the project then turned to determining the temperature distribution in the glass to metal seal area during normal use. There are many variables that will effect this temperature distribution including, solar radiation intensity and angle, ambient temperature, wind, flow rate of the heat transfer fluid, and more. Since we were not able to collect test data for all conditions it was decided to collect test data whenever it was practical and to use this test data to calibrate heat transfer models that could then be used to predict the joint temperatures under various conditions.

EES Model

This model uses a resistance model to account for the heat transfer radially through the HCE. A schematic representation of this model is shown in Figure 11. Convection between the steel and the heat transfer fluid is given by R_{12}, Radiation between the steel and the glass is R_{23}, R_{34} is conduction through the insulation, and R_{45} is combined convection and radiation out from the surface of the insulation.
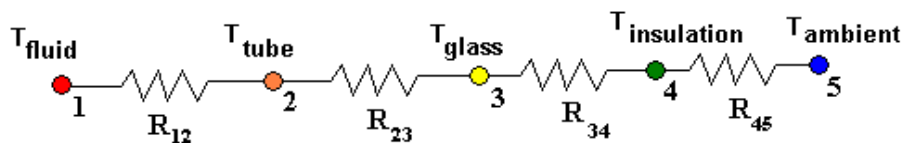
Figure 11, Resistance Model

7

The model developed here ignored end effects and heat transfer theories for infinitely long concentric cylinders were applied. EES (Engineering Equation Solver) software was used to solve the system of equations presented below to find the steady state solution.

The material properties that were important for this analysis included the transmissivity, reflectivity, and absorptivity of the glass and steel in the collection element. The glass properties were taken from data sheets provided by Corning for Borosilicate Pyrex. The steel properties were estimated as the selective coating used to maximize the thermal gain on the steel tube is proprietary and limited technical information was available. The properties used for this analysis are listed in Table 2.

Table 2, Properties Used for Heat Transfer Analysis

	Transmissivity τ	Reflectivity ρ	Absorptivity α	Emmissivity ε	Conductivity k [W/(m ^{o}K)]
Glass	0.9	0.04	0.06	0.9	1.35
Steel (selective Coating)	0	0.04	0.96	0.07	45

The model used here is based on the conservation of energy as given in   Equation   1. This concept is explained graphically in Figure 2 with E_{in} the energy into the system, E_g the energy generated in the system, E_{out} the energy leaving the system and E_{st} the energy stored in the system. In words this means the time rate at which energy enters the control volume plus the rate at which energy is generated inside the control volume minus the rate at which energy leaves the system equals to the rate at which energy is changed inside the control volume.

$$\frac{dE_{in}}{dt} + \frac{dE_g}{dt} - \frac{dE_{out}}{dt} = \frac{dE_{st}}{dt}$$ **Equation 1**
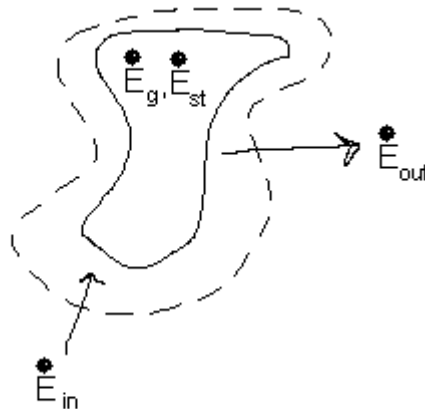
Figure 12, Conservation of Energy Model

There are two ways to consider the energy balance in the glass. One is to consider the energy coming in, q_{solar}, and the energy coming out, q_{s2out}, but this neglects any reflected radiation. The other way is to consider what is being absorbed to be heat generation in the glass. The latter method simplifies the analysis. Heat can also enter the control volume by radiation from the steel tube, q_{12}. This makes the only two ways heat can enter the glass by q_{abs}, the absorbed solar radiation, and by q_{12}. The only two methods for heat to leave the glass are through convection or by radiation. The conservation of energy for this part of the glass is represented in Figure 13.
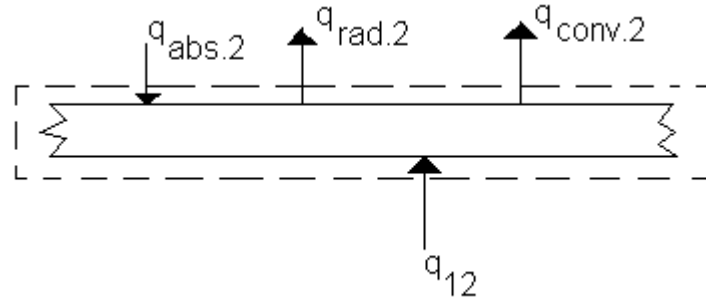
Figure 13, Heat Flows Into and Out of the Glass

This creates an energy balance as shown in Equation 2.

$$q_{12} + q_{abs2} = q_{conv2} + q_{rad2} \qquad \textbf{Equation 2}$$

Finally, a conservation of energy analysis must be done on the steel tube. q_{solar1} quantifies the amount of heat absorbed by the selective coating and thus into the steel tube. q_{12} quantifies the amount of heat radiated to the glass from the steel, thus removing heat. q_{conv1} removes heat from the steel and puts it in the heat transfer fluid. These modes of heat transfer are illustrated in Figure 14.
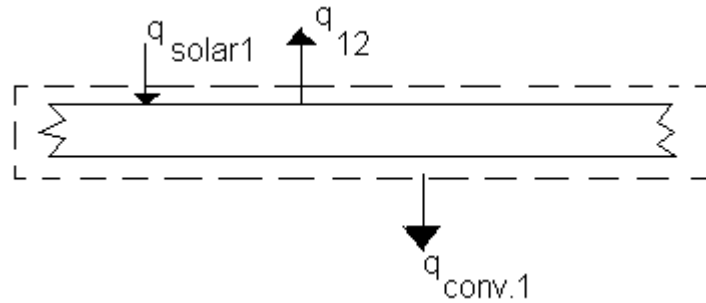
Figure 14 , Conservation of Energy Control Volume in HCE steel tube

Applying the laws of conservation of energy this produces the energy balance of Equation 3.

$$q_{solar1} = q_{12} + q_{conv1} \qquad \textbf{Equation 3}$$

The entire system is depicted in Figure 15. The inner tube is designated by surface 1o on the outside and surface 1i on the inside. Similarly the glass is designated surface 2i on the inside and 2o on the outside.
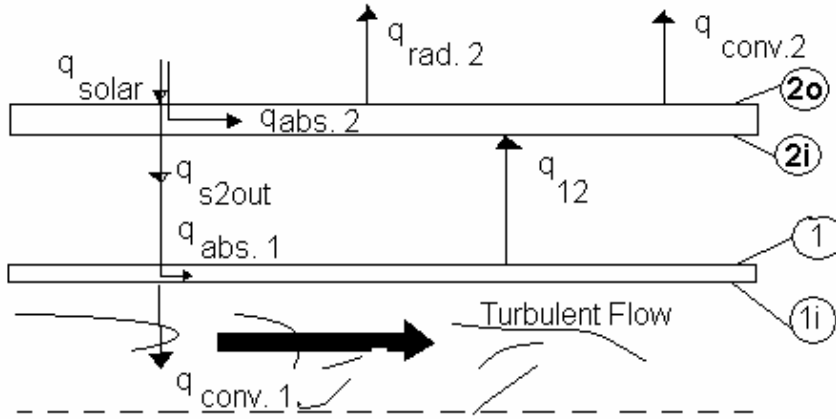
9

Figure 15, Nomenclature for HCE Heat Transfer Modes

An estimate for the incident solar radiation was provided by KJC. Normal incident pyranometers were used to monitor this data both at individual SEGs and overall at a weather station. Various efficiency parameters decrease the theoretical maximum power input to the system from the sun. These include the reflectivity of the reflector panels, the shape of the parabolic mirrors, and dirt on the reflective panels. The glass itself has some reflectivity, which again decreases efficiency. The selective coating's absorptivity is not unity and thus another loss is added. This produces an overall optical efficiency of about 76 % for the LS2 collectors. Table 3 gives a breakdown of the losses.

The solar radiation in W/m^2 is multiplied by the area of the reflective panels and the loss factors associated with the reflective panels to get the heat transfer into the HCE, q_{solar}. The reflectivity of the glass is then used to calculate the percentage of the remaining energy that does not penetrate the outer surface of the glass. Of this energy a portion is absorbed into the glass and the remaining portion passes through to the steel inner tube. A portion of the heat reaching the inner tube is radiated back towards the glass and the remainder is absorbed and transferred to the heat transfer fluid.

The amount of energy passing through the glass, q_{s2out}, is calculated using the transnissivity of the glass as shown in Equation 4.

$$q_{s2out} = \tau_g q_{solar} \qquad \textbf{Equation 4}$$

The portion of the remaining heat absorbed by the steel tube is determined by the absorptivity of the selective coating on the steel as seen in Equation 5.

$$q_{abs1} = q_{s2out} \alpha_s \qquad \textbf{Equation 5}$$

10

Table 3, Losses for the LS2 Collector

	(LS2) efficiency
Reflector panel reflectivity	94 %
Parabolic reflectivity	94 %
Cleanness of reflector panels	94 %
Anti-reflective glass coating	96.5 %
Absorptivity of selective coating	95 %
Overall	76 %

The heat transfer between the steel inner tube and the heat transfer fluid is governed by convection. Equation 6 gives the heat transfer. The temperature on the inner surface of the steel is given by T_{1i} and the temperature of the fluid is given by T_f. Heat transfer by convection depends on the convection coefficient, h_1, and the Area of contact, A_{1i}. The convection coefficient is a function of the Nusselt number of the tube, which in turn is dependent on the Pandtl number of the fluid, and the Reynolds number of the flow. The velocity of the fluid can be calculated from the volumetric flow rate and cross sectional area of the flow. This along with the properties of the fluid gives the Reynolds number. The Nusselt number and Prandtl number of the fluid come directly from the fluid properties. The name of the heat transfer fluid is Therminol. Required properties of Therminol are listed in Table 4.

$$q_{conv1} = h_1 A_{1i} (T_{1i} - T_f)$$ **Equation 6**

Table 4, Thermal Properties of Therminol at 400° C

Density ρ_f [kg/m^3]	Viscosity μ_f [Pa s]	Specific Heat c_f [J/(kg ^{o}K)]	Conductivity k_f [W/(m ^{o}K)]
703	0.17e-3	2600	0.0771

The required calculations for the convection coefficient are listed in Equations 7.

$$\alpha_f = \frac{k_f}{\rho_f * c_f}$$

11

$$Nu_f = \frac{\mu_f}{\rho_f}$$

$$\mathrm{Re}_f = \frac{\rho_f V_f d_{1i}}{\mu_f}$$

$$\mathrm{Pr}_f = \frac{Nu_f}{\alpha_f} \qquad \textbf{Equations 7}$$

$$Nu_t = 0.023 * \mathrm{Re}_f^{4/5} * \mathrm{Pr}_f^{1/3}$$

$$h_1 = \frac{k_f * Nu_t}{d_{1i}}$$

The heat absorbed by the steel is determined by the absorptivity of the selective coating.

Radiation is the only mode of heat transfer between the steel and the glass as the system is evacuated. Equation 8 defines the amount of heat transferred between these two surfaces, with σ, the Stefan-Boltzmann constant (5.67e10 -8 W/[m2 oK4]), ε, a constant based on the emissivites of the glass and the steel for radiation between two concentric tubes. A_{1o} and A_{2i} are the respective areas of the outside of the tube and the inside of the glass. The temperatures corresponding to these surfaces are given by T_{1o} and T_{2i}.

$$q_{12} = \sigma\varepsilon\left(A_{1o}T_{1o}^4 - A_{2i}T_{2i}^4\right) \qquad \textbf{Equation 8}$$

Finally the heat transfer from the glass to the surrounding environment is considered. This consists of both convection and radiation. The convection component is determined by using an analytical solution for a cylinder in a cross flow. The convection heat transfer is calculated just as for the convection to the heat transfer fluid (see Equation 6) with the temperatures for the out side of the glass and the ambient temperature of the surroundings. The convection coefficient is calculated from standard cross flow analysis. The radiation to the surroundings from the glass surface is given by Equation 9, with T_∞ the ambient temperature of the surroundings.

$$q_{rad2} = A_{2o}\varepsilon_g\sigma\left(T_{2o}^4 - T_\infty^4\right) \qquad \textbf{Equation 9}$$

It remains to determine the temperature gradients in both the steel and the glass. The steel has a high thermal conductivity and no heat generation. Thus the steel is assumed to have a very small gradient and this gradient is assumed to be linear. The glass has a relatively low thermal conductivity and a fair amount of heat generation (modeling the absorbed solar radiation). For these conditions the gradient will not necessarily be linear or small.

12

Above, a thermal resistance method is introduced that roughly states that $q = \Delta T/R$, with q the heat flow through the material, ΔT the temperature difference across the material, and R the thermal resistance of the material. Assuming a linear temperature distribution and applying this equation we obtain the following for the steel tube:

$$q_{s2out} = \frac{(T_{1o} - T_{1i})}{R_{tube}} \qquad \textbf{Equation 10}$$

R_{tube} is the thermal resistance of the steel tube. This thermal resistance is given by:

$$R_{tube} = \ln\left(\frac{r_{1o}}{r_{1i}}\right) 2\pi L_{tube} k_s \qquad \textbf{Equation 11}$$

With L_{tube} the length of the tube. A thorough analysis is presented in [1].

For the glass tube a slightly more complicated solution is involved. This is because the solution depends on more complicated heat fluxes and boundary conditions. First, an approximate calculation of the heat generation must be made. Assuming the heat is generated uniformly simplifies the problem. The volumetric heat generation term, $\dot{g}$, is then given by:

$$\dot{g} = \frac{q_{abs}}{Vol_g} \qquad \textbf{Equation 12}$$

The analysis is similar to that which is done in [1]. This general solution is:

$$T(r) = \left(\frac{-\dot{g}}{4k_g}\right) r^2 + C_1 \ln r + C_2 \qquad \textbf{Equation 13}$$

The radial coordinate through the glass in given by r and C_1 and C_2 are constants that result from integration. The constants must be solved for via boundary conditions. Upon differentiating the general form the following equation is produced:

$$\frac{dT}{dr} = \left(\frac{-\dot{g}}{2k_g}\right) r + \frac{C_1}{r} \qquad \textbf{Equation 14}$$

Applying this equation at the outer and inner surfaces of the glass gives the following 2 equations respectively:

13

$$\left(\frac{dT}{dr}\right)_{r=r_{2o}} = \left(\frac{-g}{2k_g}\right)r_{20} + \frac{C_1}{r_{2o}} \quad \textbf{Equation 15}$$

$$\left(\frac{dT}{dr}\right)_{r=r_{2i}} = \left(\frac{-g}{2k_g}\right)r_{2i} + \frac{C_1}{r_{2i}} \quad \textbf{Equation 16}$$

Another equation involving the temperature gradient at the inside surface can be written by setting the conduction into the surface equal to the radiation coming in from the steel tube.

$$\left(\frac{dT}{dr}\right)_{r=r_{2i}} = \left(\frac{q_{12}}{-k_g A_{2i}}\right) \quad \textbf{Equation 17}$$

The heat flux at the outer surface of the glass is given by the following equation:

$$q''_{out} = \left(\frac{q_{rad2} + q_{conv2}}{A_{2o}}\right) = -k\left(\frac{dT}{dr}\right)_{r=r_{2o}} \quad \textbf{Equation 18}$$

This equation can provide a check to see whether the heat fluxes and boundary conditions were applied properly.

This model was used to help calibrate the finite element model that is presented in the next section.

Finite Element Model

The next step in the heat transfer analysis was to construct a three dimensional finite element model of the HCE joint area. The goal of this analysis was to predict the temperature and temperature gradients in the region around the glass to metal seal. This model allows for the inclusion of boundary effects including the conduction from the steel tube through the expansion bellows into the glass to metal seal. The model was calibrated to test data and then used to predict worst-case temperatures for lower sun angles. The solid model used to create the finite element model is shown in Figure 16. The complex geometry of the expansion bellows was not actually carried over into the finite element model as it is too complex and makes meshing very difficult. Instead a simplified model was used as shown in Figure 17.

14
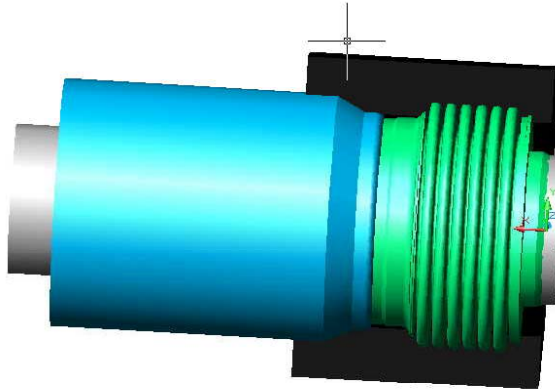
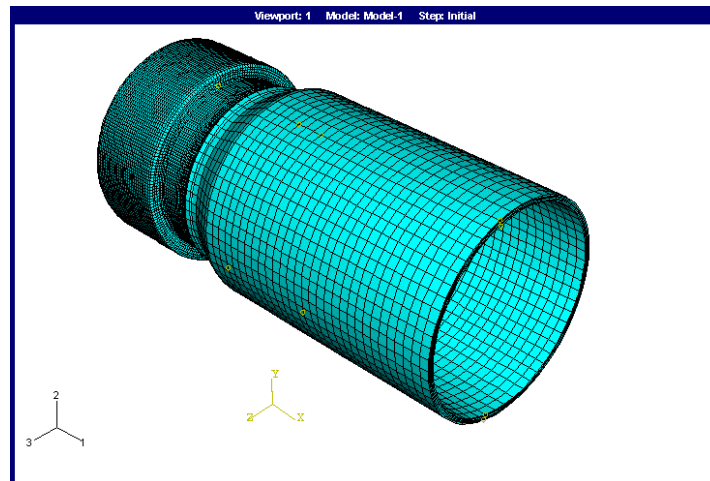**Figure 16, Solid Model Used of HCE**

Figure 17, Simplified Model for Finite Element Analysis

The mesh for this model was determined to be adequate partly based on the results from the EES model presented earlier. The temperature distributions through the thickness of the glass were seen to be fairly linear. The final model used 5 linear brick elements through the thickness of the glass and 2 linear brick elements through the thickness of the steel collar.

Much thought and labor went into creating a model that would converge to a reasonable solution without taking an inordinate amount of time to process. Radiation between the glass and the steel tube presented a major challenge as the computational power required to do the calculations properly with the complicated geometry of this joint was not available. Approximate solutions to this problem were found and the model correlated fairly well to the test data collected.

Details on the execution of this finite element analysis are to be reported elsewhere by the co-author in partial completion of his masters degree in mechanical engineering [2].

15

Figure 18 shows the temperature distribution as predicted by the model with inputs matching the test days as closely as possible. This model attempts to predict the steady state temperature in the flatter part of the temperature versus time curve as reported in the test section.
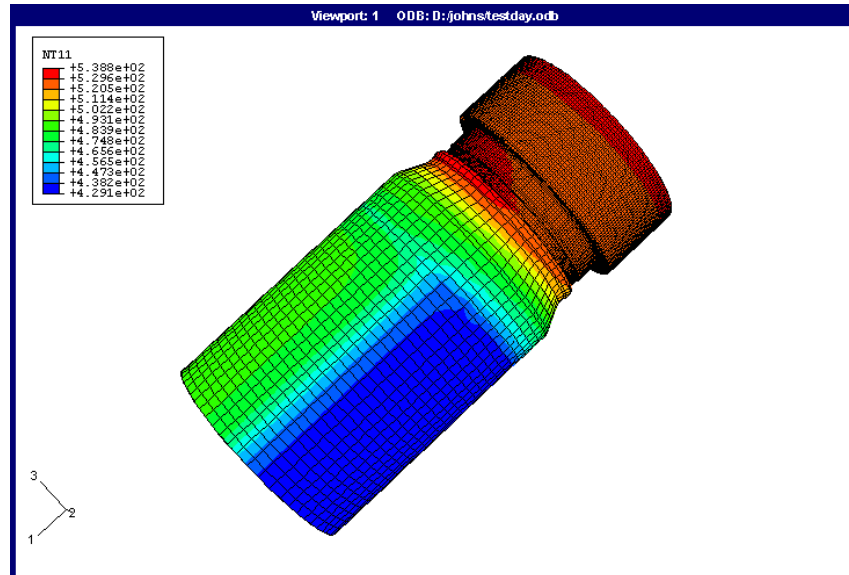
Figure 18, Predicted Temperature Distribution for Test Days

The model parameters were then changed to simulate the lower sun angles encountered in the spring. The results of this analysis are shown in Figure 19.
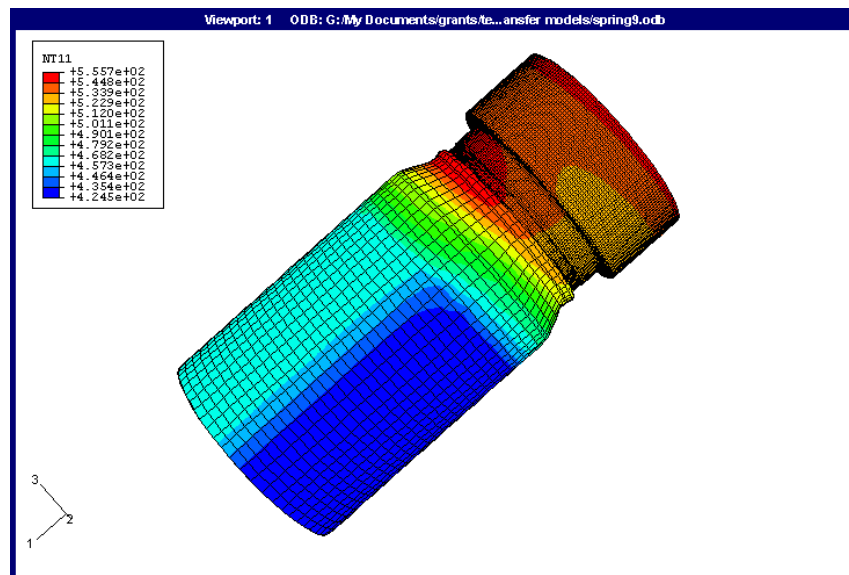

Figure 19, Predicted Temperature Distribution for Spring Sun Angles

16

The addition of the radiation reflecting into the joint is predicted to raise the peak temperature of the joint by approximately 17 °C. Also it can be seen that the temperature gradients around the joint are more severe for this case.

Correlation between the finite element models and the test data are discussed in the next section.

Correlation of Models to Test Data

The ABAQUS models did a fairly good job of predicting the temperatures in the joint area for the test days. Figure 20 shows a comparison of the predicted to actual temperatures at 15:00 on August 29th. The discrepancy at position 6 on the glass is not very important to this analysis and is greatly affected by the current wind conditions as shown above in Figure 10. The circumferential temperature distribution around the seal correlated well between the model and the test data. The model showed a temperature range of 12 °C circumferentially at the seal (position 5 of the test data) which was very close to the variation observed in the test data.
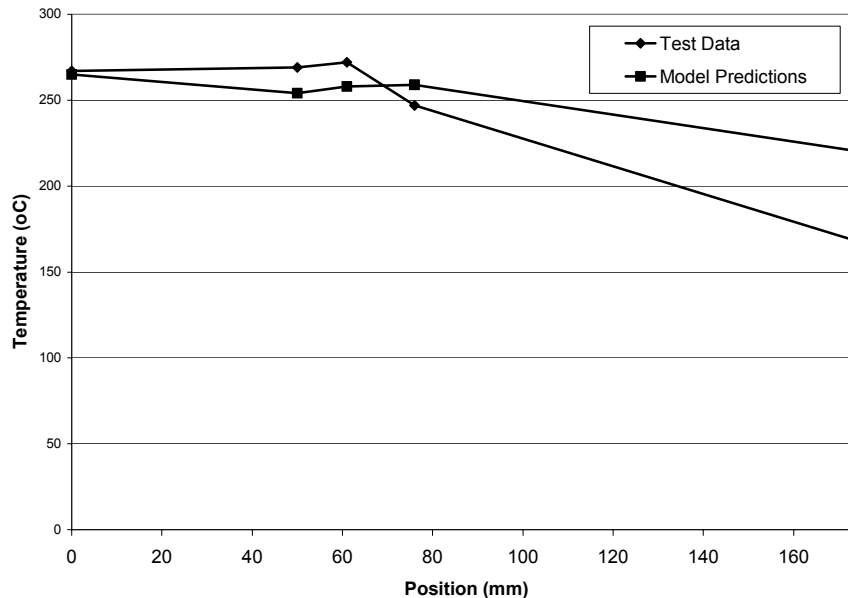
Figure 20, Comparison of ABAQUS Predicted Temperatures to Test Day Temperatures

Analysis of Existing Configuration

The existing glass to metal seal was investigated using finite element analysis. The models used to perform this analysis are predominantly meant to highlight problem areas in the joint. A rigorous finite element model of this joint requires that the full nonlinear behavior of the stainless steel sleeve and glass be accounted for and that the circumferential temperature variation also be taken into consideration. Some of the nonlinearities of these materials have been included such as the yield strength, σ_{yield}, of the steel, the temperature dependence of the steel and glass moduli, E, and Poisson's ratios, ν, and the temperature dependence of the coefficient of thermal expansion, α, for

17

the steel (the coefficient for the glass is fairly constant in the range of interest). The brittle fracture of the glass is included but the models at this stage are fairly crude. The results from these models are used more qualitatively than quantitatively to gain insight into the joint behavior. Table 5 shows the properties assumed for the stainless steel collar. Table 6 gives similar properties for the glass.

All of the models presented in this section assume that the joint can be analyzed as an axisymmetric system. This assumption requires the temperatures, loads, and boundary conditions to be the same for any point chosen around the circumference of the joint. This is not strictly a valid assumption as the test data does show temperature variation around the circumference. This variation was deemed low compared to the absolute temperature and is therefore ignored throughout this preliminary analysis.

Table 5, Material Properties used for Stainless Steel Collar

T (^{o}C)	E (GPa)	ν	α (10^{-6} m/m/^{o}C)	σ_{yield} (MPa)
21	202.7	0.26	17.73	44.66
38	199.3	0.27	17.91	41.22
93	195.8	0.28	18.23	39.51
149	192.4	0.30	18.63	34.35
204	185.5	0.31	18.88	32.64
260	178.6	0.32	19.08	34.35
316	171.8	0.33	19.35	36.07
371	168.3	0.33	19.44	44.66
427	164.9	0.33	19.78	54.97
482	158.0	0.33	19.98	65.27
538	151.2	0.32	20.21	68.71

Table 6, Material Properties Used for Borosilicate Glass

T (^{o}C)	E (GPa)	ν	σ_{crack} (Mpa)
0	62.90	0.20	6.9
100	63.50	0.20	6.9
200	64.00	0.20	6.9
300	64.50	0.19	6.9
400	65.00	0.19	27.25
500	65.50	0.19	27.50
600	64.50	0.22	26.50

*Note:α is approximately constant at 3.25 e -6 m/(m/^{o}C)

Details of the manufacturing process are not available at this time but it is difficult to imagine that the joint will be stress free at room temperature. The viscoelastic nature of the glass may relieve some of the residual stresses left over from the cooling process but

18

it is inevitable that some stresses will remain. These residual stresses if present would more than likely be beneficial during the subsequent heating of the joint. The current analysis does not assume that any residual stresses are present in the joint at an ambient temperature of 25 $^\circ$C.

A linearly increasing temperature is applied to the joint starting at the ambient temperature. The maximum principal stresses in the joint are monitored for the glass to determine were cracking is likely to occur. The series of frames depicted in Figure 21 show the progression of the maximum principal stress in the joint as the temperature is increased. The red areas represent areas of high tensile principal stresses. These figures clearly indicate that the portion of the glass bulb inside the HCE experiences the highest stresses and is most likely to crack. This correlates well with the experience at KJC where small pieces of glass are often evident inside the HCE as seen in Figure 22. Figure 21d) shows that eventually the entire inner bulb may crack away from the rest of the glass as indicated by the red zone on the outer neck of the bulb. This may indicate the ultimate failure mode of the HCE. As mentioned above these models do take into account cracking. This is not an easy phenomenon to visualize as a given region still carries stresses after cracking just not in the direction of the crack. Inspection of XX shows areas that exhibit cracking early on show low stress later. This is due the cracking of the material in those areas. The maximum principal stress can increase after cracking due to the orientation of the principal stresses shifting. If a region cracks in three orthogonal directions then the ability to carry tensile load is lost for that region. Again it must be emphasized that the results reported here are simply for evaluation of the configurations relative to each other. The actual predicted temperature at failure would be highly dependent on the residual stresses in the joint as well as relaxation effects.
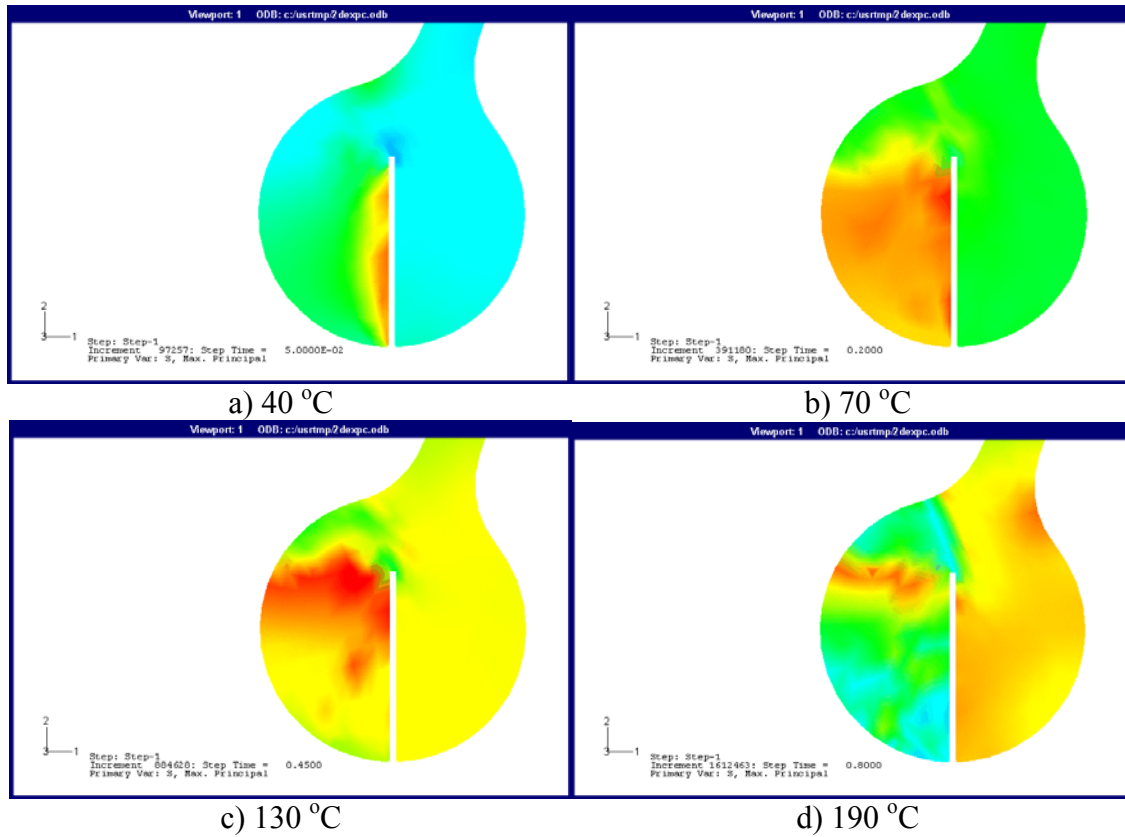
19

| a) 40 $^{\circ}$C | b) 70 $^{\circ}$C |

| c) 130 $^{\circ}$C | d) 190 $^{\circ}$C |

Figure 21, Principal Stresses in the Existing Configuration as Temperature is Increased
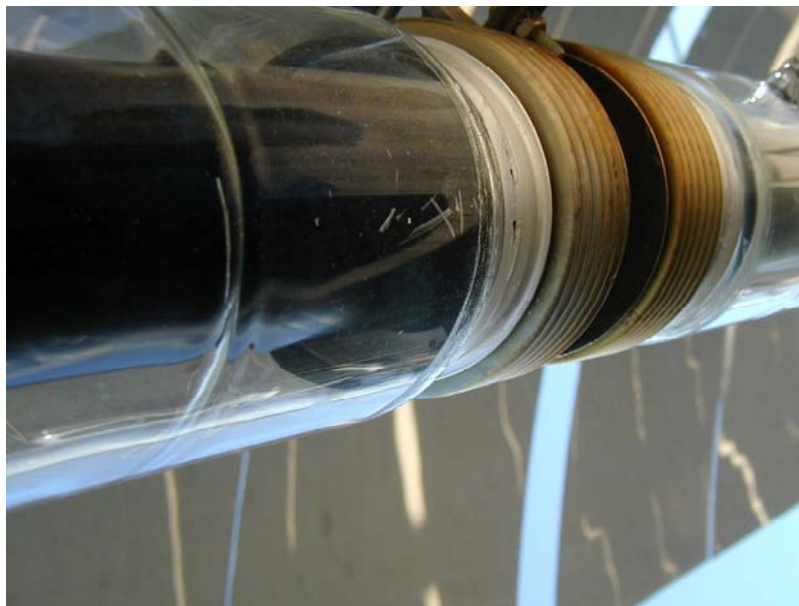

Figure 22, Glass Shards Evident Inside of The HCE

20

Potential Solutions

Modifying the Material of the Collar

A relatively simple fix for this application would be to find a material that has a coefficient of thermal expansion closer to that of the glass that can still meet all of the other required specifications for this design. Some of these requirements include the ability to be attached to the expansion bellows, the ability to take the high temperatures of the joint, the material must be corrosion resistant, and must be able to maintain a vacuum. Several candidates have been investigated preliminarily for this application.

Titanium has a coefficient of thermal expansion approximately 55%-65% that of the stainless steel in current use. An alloy such as Ti-35A would give lower strength than the stainless in the current joint. This may be beneficial to limit the stresses in the glass. If higher strength is required Ti-50A or even a high strength alloy such as Ti-6Al-4V could be used. In all cases the lower modulus of elasticity of the titanium (just more than half of the stainless) would be beneficial to lowering the stresses in the glass. Table 7 lists the pertinent material properties for some titanium alloys. Figure 23shows the maximum principal stress with a titanium collar. Ti-35A was assumed for this analysis. It is believed that an appropriate alloy could be found that could still be welded to stainless steel. This operation may not be trivial however.

Table 7, Mechanical Properties of Titanium Alloys

	Modulus (Gpa)	Coeficient of Thermal Expansion $(x10^{-6} /^oC)$	Yield Strength (MPa)	Ultimate Strength (MPa)
Ti-35A	114	9	172	241
Ti-50A	114	9	276	345
Ti-6Al-4V	114	9	850	900

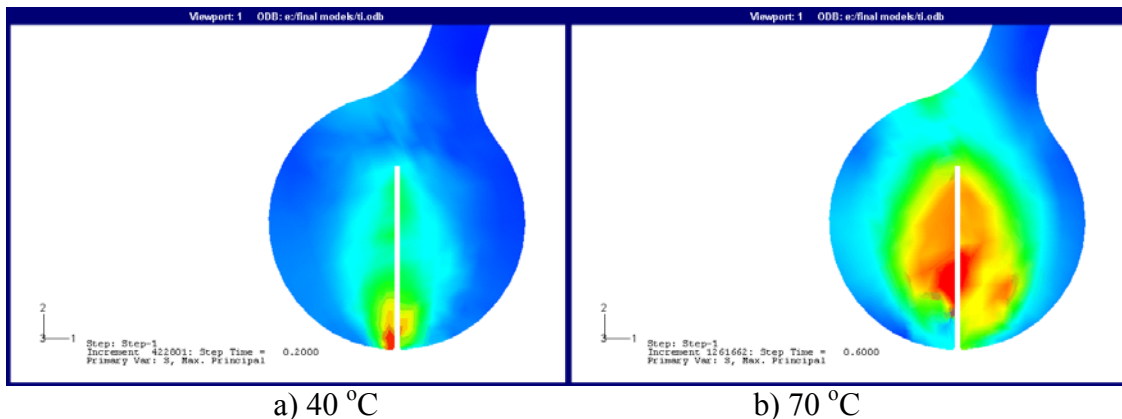a) 40 ^{o}C   b) 70 ^{o}C

Figure 23, Maximum Principal Stresses in a Titanium Joint

The principal stress distribution in the titanium joint is much better distributed than in the steel joint. At 70 ^{0}C the inner section of the glass bulb has largely cracked out with the steel collar but still has quite a bit of strength with the titanium collar.

21

102

Another material being investigated at this time is Invar. Invar is a nickel alloy with a low coefficient of thermal expansion. Typical properties of Invar 36 are listed in Table 8. Other low expansion alloys should also be investigated further for this application. Cost may be an issue for some of these alloys as will be connection to the expansion bellows.

Table 8, Mechanical Properties of Invar 36

	Modulus (Gpa)	Coeficient of Thermal Expansion $(x10^{-6}/^{o}C)$	Yield Strength (MPa)	Ultimate Strength (MPa)
Invar 36	150	6*	300	450

*Varies greatly with temperature

Advanced Composite Solutions

Advanced composite materials are composed of stiff fibers embedded in a relatively soft matrix. These materials offer unique opportunities in a design application such as the HCE joint. Many of the material properties such as stiffness and thermal expansion coefficient can be tailored by varying the fiber orientations in the material. The designs presented below take advantage of the extremely low coefficient of thermal expansion that can be achieved with proper choice of constituents for a composite material. If indeed the HCE is breaking due to the differential expansion between the glass and the stainless steel connector then perhaps a third component can be added to the joint to minimize the stresses seen in the glass by limiting the expansion of the steel. Again the high temperatures of this application will make finding an appropriate material difficult. Carbon fibers are an ideal candidate for the reinforcement as they have a slightly negative coefficient of thermal expansion, which leads to a composite material when combined with a matrix with a very low positive coefficient. The matrix material will be more problematic as the high operating temperatures may rule out many of the commonly used options.

The preliminary analysis performed here has not made any attempt to vary the material properties of the composite to evaluate the effect on the joint. The wraps used here are composed of uniaxial carbon fibers wrapped in the circumferential direction. The assumed properties of the wrap are given in Table 9. Note that this is not an isotropic material. The 2 direction is circumferential.

Several different scenarios utilizing advanced composites have been evaluated. The first is a potential retrofit for existing HCEs in service. The concept for this design is depicted in Figure 24. This design calls for a composite wrap around the joint area. The wrap could extend up over the glass bulb as shown or be terminated below the glass only wrapping the steel sleeve.

22

Table 9 Material Properties for Composite Wrap

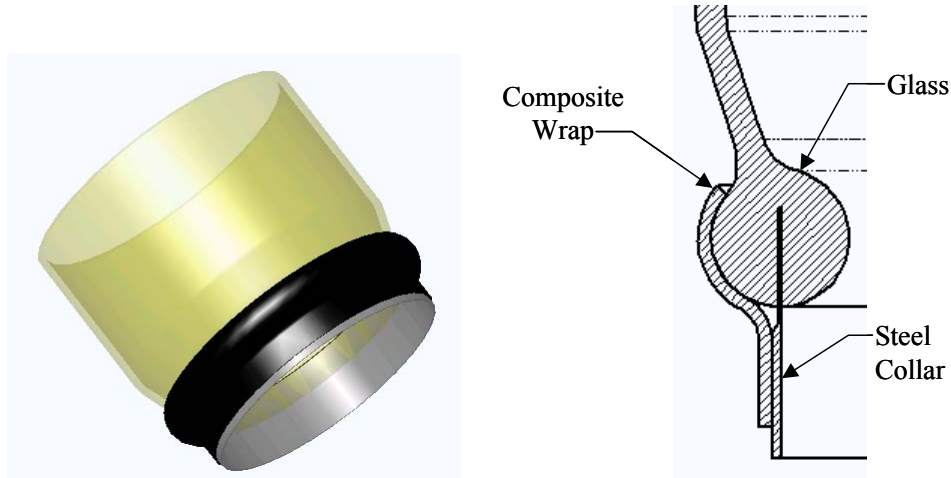| E_1 (GPa)	E_2 (GPa)	E_3 (GPa)	υ_{12}	υ_{13}	υ_{23}	G_{12} (GPa)	G_{13} (GPa)	G_{23} (GPa)
10.9	159	10.9	0.026	0.4	0.38	6.4	1.0	6.4

Figure 24, Composite Wrap Design

The second option that was investigated in this area was similar to the first but placed the wrap on the collar before it was embedded in the glass. Obviously this solution is no longer a retrofit and only applies to new production. This concept is depicted in Figure 25.
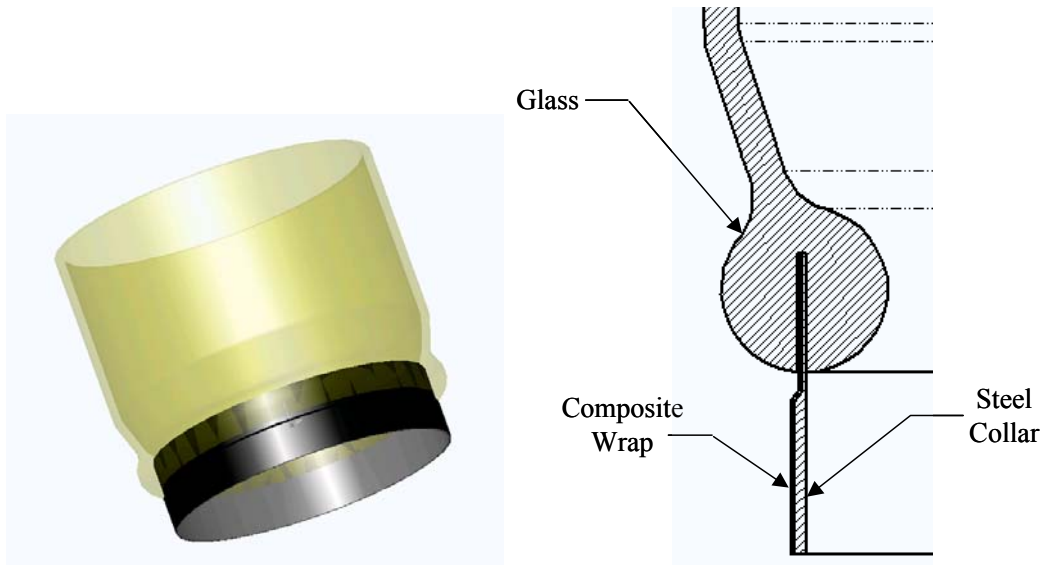
Figure 25, Internal Composite Wrap

The maximum principal stresses for the external and internal wrap are shown below in Figure 26.

a) External Wrap   b) Internal Wrap
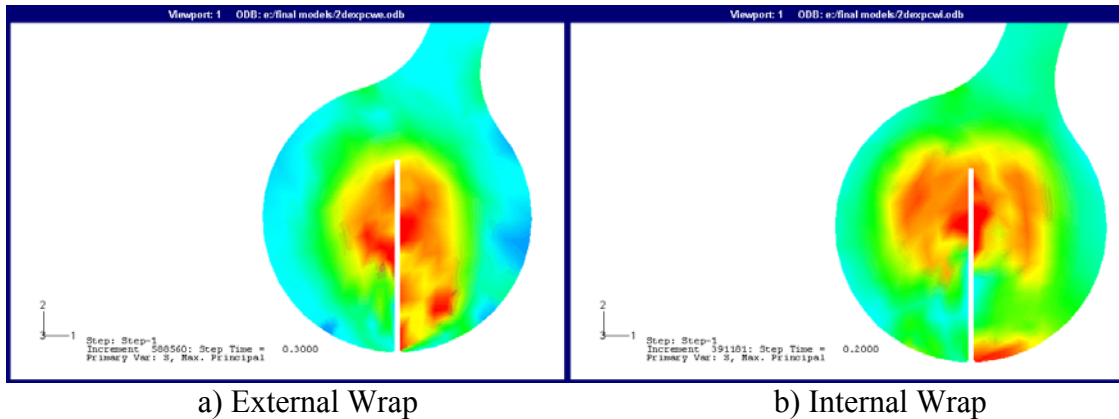
Figure 26, Maximum Principal Stress with Composite Wrap

The results in Figure 26 indicate that the composite wrap can be effective in altering the distribution of maximum principal stress in the joint. The stress distributions seen for the wrap solutions both tend to concentrate the cracking of the glass to a narrower region around the steel interface than the original configuration (see Figure 21). The wraps used here are very effective in limiting the radial expansion of the steel collar but due to the chosen fiber orientation and distribution they do not stop the steel collar from wanting to pull out of the glass bulb as it expands. The internal wrap solution could theoretically help with this mode if some fibers were oriented along the long axis of the tube. Such a configuration may work to limit the steel collar expansion in all directions. The results thus far are inconclusive as to the applicability of this solution to the actual joint. More study on composite configuration and manufacturing issues would need to be addressed to assess the overall suitability of these design alternatives.

---

[1] Incorpora and Dewitt, <u>Fundamentals of Heat and Mass Transfer</u>, John Wiley and Sons, New York
[2] John Myron Allen, <u>Heat Transfer Analysis of a Heat Collection Element,</u> Masters Thesis, Cal Poly, San Luis Obispo, CA

24

Range Sensing and Real-Time Registration 2

Fred W. DePiero

Electrical Engineering Department
Cal Poly State University

Final Report
Project: 54530
Title: Range Sensing and Real-Time Registration 2
Sponsor: Office of Naval Research (ONR)/N00014-02-1-0754

The application area for my work is an advanced type of remote viewing system. The system would provide greater flexibility than with conventional methods, such as are used in teleconferencing. Using my techniques each user could select their own viewpoint of a remote scene. However this is accomplished without installing a separate camera (and camera positioner) for each user – such an approach would not scale well for multiple users. Rather my method combines sensor data from multiple perspectives and then renders (computes) views for each user. The challenges of my work are accuracy when combining sensor data (registration) and the development of methods that require deterministic worst-case processing times. These are both important for real-time viewing. Sensor positioning and motion planning are also important to mitigate occlusion effects when rendering imagery.

This past year I improved my registration algorithms and worked on a networked implementation that includes transmission and visualization components. I collaborated with other faculty and with students. Professor Chris Buckalew of the Computer Science Department is a computer graphics expert and I found his input and suggestions quite helpful. I also worked with several graduate students. Tim Jackson is working on a thesis directly related to this project, "Stable and Accurate Landmark Extraction in Range Images". Kurtis Kredo, Ryan Manes and Brian Gleason worked on communication and control software for a four-processor computer system. This networked system was designed to support a pipelined implementation of processing algorithms. This improves processing throughput, while still using a standard computer platform.

With the inclusion of the networked transmission and visualization components as well as the algorithmic improvements, I now have a near complete, end-to-end implementation of a remote-viewing system. My setup does lack a real-time sensor ($100k) but development of the processing components is well established.

The baseline techniques that I am developing (registration) also have a number of other potential application areas, such as the automatic construction of digital elevation maps. This is of interest to potential military sponsors. Pursuing the remote viewing system seems to be a fruitful route for the present. This type of system seems to resonate well with sponsors, who seem to appreciate potential spin-off applications.

The following publications describe technical details of my work this past year. Both of these papers appeared in peer-reviewed, international conferences.

F. W. DePiero, "Deterministic Surface Registration at 10Hz Based on Landmark Graphs With Prediction," *14th British Machine Vision Conf. (BMVC2003), Norwich, UK, Sept 9-11, 2003.*

F. W. DePiero, "Surface Registration at 10Hz Based on Landmark Graphs - Benefits for a Scalable Remote Viewing System," *3rd International Conf on Visualization, Imaging, and Image Processing (VIIP2003), IASTED, Benalmadena, Spain, Sept 8-10, 2003.*

Deterministic Surface Registration at 10Hz
Based on Landmark Graphs – With Prediction

Abstract

Landmark graphs provide a means for surface registration, based on determining subgraph isomorphism to find scene-to-scene correspondences. Surface data used herein included both range and colour imagery. Images were acquired of a static scene from a moving sensor. The continuous motion allowed the sensor position to be predicted. This stabilized graph formation. Landmarks were determined using the KLT corner detector. Graph structure was established using nodes (landmarks) and edges that agreed well with predicted locations. Subgraph matching was approximated using the LeRP algorithm. A 6 DOF rigid transformation including translation and rotation was found via Horn's method. Test results on real and synthetic images indicate that a substantial speed improvement is possible, with greater determinism than ICP, while maintaining accuracy. Tests incorporated relatively large image displacements, spanning up to 30% of the sensor FOV for the image stream. Mean absolute errors of 0.6-0.9% FOV were found. Mean compute rates were ~10 Hz with standard deviation ranging 6-9%, for an image size of 200x200. Tests were run on a 900 MHz PC. 32 test trials are reported, with comparisons against a fast version of ICP.

1 Introduction

Future viewing systems could permit multiple users to explore scenes of remote locations with greater flexibility than is possible today. Rather that provide each user with their own video camera, positioner and transmission channel, the techniques described herein support a different approach. Images for each user would be rendered from a set of 3-D surface data, based on a selected viewpoint. Surface registration is the process of aligning the 3-D data into an expansive, contiguous set – beyond the field of view of a single sensor image.

Tele-presence viewing and immersion VR require 3-D sensing, registration, transmission and visualization technologies, some of which are unavailable today for real-time systems. For example a system using small 320x200 sensor images acquired at 25 Hz (motion picture rate), would need to acquire and process nearly $2 \cdot 10^6$ points/sec. For perspective, some state-of-the-art commercial 3-D sensors are within an order of magnitude of this rate – such as the Perceptron LASAR at 325k pts/sec and others [31]. Real-time registration is another key technology, in addition to sensing, that is very demanding and still requires further study. A rate of 400k pts/sec is reported herein.

Given a sensor that acquires both surface location and colour, and given a real-time registration capability, it would become possible to render live images from viewpoints that are distinct from sensor locations. This would provide a substantial degree of flexibility for remote viewing systems compared to current systems. Applications such as a 'television with a joystick' would become possible. For example in a sports broadcast, some viewers might choose to watch the hands of a golfer, others the ball, others the whole putting green. Granted more complex scenes might have a high level of occlusion, despite multiple sensors.

For tele-immersion, two such views could be computed, one for each eye. Another application is tele-medicine, where this sort of system could provide useful flexibility. For example if a field technician positioned a range & colour sensor over a patient's wound, then a remote doctor could examine the injury. Furthermore, if the doctor's viewpoint were graphically presented to the sensor technician, then the technician could anticipate the doctor's viewing needs – in terms of standoff or locations, for example. This would permit a much more efficient viewing experience for the doctor. In

another approach, the technician could possibly be replaced by a robot, which would use the doctor's viewpoint as a basis for path planning when positioning the sensor.

All of these advanced viewing systems rely on surface registration that can be accomplished in a fast and deterministic fashion. The fundamental reason that registration is required is because sensors such as laser range finders (or simple video cameras) are all line-of-sight devices. Hence either multiple sensors or multiple images (from a moving sensor) would typically be required to form a complete set of surface data across an entire scene. The focus of this effort has been on static scenes with a moving sensor.

2 Approach

The long-term goal of this research is to pursue a technique that performs view registration automatically at rates ~25 Hz, with relatively large image sizes and large sensor displacements. This performance goal targets remote viewing applications, with rapid sensor motion.

To achieve fast and deterministic processing, iterative [1] [8] [13] [17] [18], compute intensive [10], or random [6] approaches were avoided. Note that reported methods often do not separate the steps of determining corresponding points and determining the transform [8]. This limits compute speed. In the new landmark-graph approach these steps have been kept separate, and are implemented in a non-iterative fashion. This is an important distinction. Another difference is that correspondence between the data sets is determined only for select landmarks as opposed to traditional ICP, for example [1], where an entire range image is used in the cost function. This improves processing speed, but it does potentially limit accuracy because not all the scene data is used to find the transform.

The landmark-graph method does not rely on an affine motion model [25][24][14]. This permits relatively large disparities to be accommodated. For example, some of the image sequences in [14] appeared to contain a displacement of ~1/5% of the FOV between subsequent images. Tests scenes documented herein ranged 2.5% to 5%. Thus the landmark-graph approach is an alternative to a stereo-based approach [29] for the large disparities.

Stability of the landmark-based registration is achieved by maintaining consistent inter-landmark (3-D) geometry, which is verified via the LeRP [4] subgraph-matching algorithm. This is in contrast to approaches such as [14] which provide robustness based on checks of deviation in the path of each individual feature, but that do not enforce a specific geometrical structure (attributed graph) between features.

A few remaining points distinguishing the landmark-graph approach from other reported methods: this method does not require photometric normalization as with [25]. It is also not reliant on a smooth surface assumption as with [2]. No particular assumptions regarding scene content (such as planar surfaces [28]) are made – but that the KLT landmark locator [12] by employable. (KLT responds to corners or sharp prominences in the range data). And finally, the method does solve for all 6 DOF of the translation and rotation, as opposed to [24][27].

2.1 Notation and Processing Summary

The following notation is used to describe the processing and representation of a stream of surface data. The stream is composed of a sequence of range and colour images, indexed by i = 0,1,2...

    I. F_i, Sensor coordinate frame for i_{th} scene.
    II.   (R_i, C_i) Range & color images acquired at location F_i.
    III. L_i, Set of landmarks found in R_i (w/rt F_i).
    IV. G_i, Graph formed from landmarks L_i.
    V. T_i, Coordinate transform relating F_i to F_0.
    VI. V_0, Graph associated with all landmarks for entire image stream.
    VII. V_i, Predicted subgraph of V_0, approximating G_i.
    VIII. V_i', and G_i' are matched subgraphs.

The following steps are used to process each (R_i, C_i) sensor input, to yield a coordinate transform T_i.

    I.   Acquire new sensor images R_i & C_i.
    II.   Predict V_i based on V_0 and on estimate of sensor motion.
    III.   Find landmarks L_i in range image R_i. Reject unstable Li.
    IV.   Form G_i using L_i, mimicking structure of Vi, (both nodes and edges).
    V.   Compute attributes for G_i, using R_i & C_i.
    VI.   Use LeRP algorithm to match G_i to V_i, the resulting subgraph
          mapping gives the L_i to L_0 correspondences.
    VII.   Find transform T_i via Horn's method, using the L_i to L_0
          correspondences.
    VIII. Check residual error from Horn. Remove outliers and recompute T_i.
    IX.   Repeat

Table 1: Notation and Processing Steps for Landmark Graph-Based Registration

2.2 Determining Landmark Location

Important performance goals for landmark detection and localization are: accuracy of location, speed and determinism of computations, and stability. Here, stability of detection refers to the consistent appearance of a given landmark, despite small shifts in sensor position, or despite sensor noise.

Figure 1. Example of scene landmarks found in two range images. Landmarks are determined via the KLT corner detector. These images contained a 5-degree shift.

Some reported techniques use invariant features that involve curvature classification, moments, or spherical harmonics [18]. These kinds of features rely on local smoothness for proper surface characterization. Jump discontinuities [19] violate this smoothness assumption. Local peaks may be adequate in some applications, such

as registering aerial range imagery [23]. However, this simple form of detection may not be sufficient to handle generic scenes.

Several methods for landmark detection were investigated in this effort. It was observed during experimental trials that ridge curves [20][21] appeared to be relatively stable. However, this approach was not pursued in light of challenges in parameterizing their shape. The use of neural nets, operating on a binary image of jump discontinuities [19] was also studied. However, the corner detector developed for the KLT algorithm (Kanade-Lucas-Tomasi) [12] appeared to be more tolerant to sensor noise, in terms of node stability and accuracy, compared to the neural networks.

KLT computes the eigenvalues associated with a covariance matrix of intensity gradients. The gradients are computed over a small 5x5 window. A 'corner' is associated with jump discontinuities that occur (ideally) along two perpendicular axes. Corners give rise to two large eigenvalues for the covariance matrix. An implementation from [7] was integrated into the system. Note that only the feature detector of [12] was used, not the tracking part of the KLT algorithm.

It may be appropriate to mention that care should be exercised when finding landmarks near jump discontinuities. A landmark should be selected on the nearer side of the jump. The scene location at the far side of the jump – along the sensor line-of-sight at the discontinuity – is inherently unstable. A ridge along a cylindrical surface is another example of an inherently unstable landmark. Using the KLT corner detector, and selecting the nearer surface, avoids these kinds of instabilities.

Post processing was used to improve the accuracy of landmark locations. This was necessary in part because the [7] implementation incorporates a multiresolution image pyramid, for improved speed. This does however appear to limit accuracy. The post processing also assured that the landmark was on the nearer surface of a jump discontinuity.

In the post-processing step, a small neighbourhood surrounding the original location (from KLT) was examined to find the sharpest corner. The nearby pixel with the largest fraction of distant neighbours was selected as the new landmark. The more distant neighbours were identified using a simple 'jump' threshold – in the range image. Hence, the measure of the local sharpness was simply the fraction of more distant neighbours surrounding the landmark. This attribute proved useful for colouring the landmarks (described next) and for eliminating unstable landmarks.

The KLT corner detector reacts to any variation in shape that occurs within a window of pixels. It is possible that within a given window, a foreground object might have a simple straight edge, while a background region might present additional curvature. The net result can make the KLT detector misfire. This is an undesirable result as the straight edge of the foreground object is not a corner, and hence is an unstable location. The local sharpness measure helped eliminate these cases. A threshold of 60% was applied to the count of more distant pixels. This eliminated the straight edge cases. This general sort of refinement of the results of KLT is also discussed in [14].

2.3 Determining Landmark Attributes

Node colours were represented by a 2x1 vector quantity. The components included hue from the colour image and the measure of local sharpness, described above.

Due to imperfect alignment of the sensor's range and colour images, it is desirable to provide improved robustness when combining data from these inputs. The examination of errors when using an inter-camera calibration model revealed problems near jump discontinuities (as to be expected). A simple approach of median filtering the hue values in the (mapped) neighbourhood of the landmark appeared beneficial.

2.4 Determining Graph Structure and Edge Attributes

Previously implementations of landmark graphs [23] used a Delaunay triangulation [5] [16] to determine which node locations were linked by an edge. For applications with

large standoffs, such as with aerial range imagery, the triangularization may be an acceptable approximation. However, the assumption of planarity was not appropriate for the close-range scenes used in this investigation.

As described in Table 1, a graph V_0 is maintained for all landmarks in the sensor data stream. As such, V_0 could potentially grow without bound – depending on the sensor trajectory. This is undesirable from the standpoint of subgraph matching. A limit on graph size is preferred for computational speed. To provide this size limit a subgraph V_i, of V_0, is predicted. The subgraph V_i contains nodes that are expected to be within the field of view of the sensor, given the sensor's previous position and the velocity estimate. V_i is formed using V_{i-1} and then adding or removing portions of V_0.

A graph G_i must also be generated for each new set of sensor data (R_i & C_i). The structure of G_i is established in a manner that mimics the structure of V_i as much as possible. For this, an edge in G_i is introduced between two landmarks L_i^a and L_i^b if

$$| L_i^a - L_0^x | < t \text{ and } | L_i^b - L_0^y | < t$$

Where L_0^x and L_0^y are a pair of landmarks in G_0 that are connected by an edge, and 't' is some appropriate threshold of distance.

This can lead to the introduction of 'redundant' or 'parallel' edges in G_i, compared to the structure of V_i. As such, graphs G_i tend to be supersets of ideal matching subgraphs. This is acceptable because the subgraph matching algorithm works well in the presence of this type of structural difference. See test trials described in [3].

The attribute, or colouring, for graph edges is simply the Euclidean distance between nodes. Distances were computed in world coordinates, not just a pixel distance, making the edge colouring tolerant to standoff changes. This formed an object-centered representation that could be compared without first aligning the range images.

2.5 Approximating Subgraph Isomorphism

Noisy sensor data introduces some fundamental limitations to the accuracy and stability of landmarks. This limits the similarity of scene graphs. Varying occlusion with different sensor viewpoints also limits the similarity of scene graphs. For these reasons, graphs made from real sensor data tend to be imperfect representations. Since the graphs are imperfect, an exact method of determining subgraph isomorphism has limited use – and consumes inappropriate compute time. Hence using an approximate method of graph matching is a reasonable approach.

Subgraph matching is accomplished using the LeRP Algorithm [4]. Also see Appendix 1. LeRP approximates a subgraph isomorphism via a deterministic procedure, based on the comparison of length-r paths. The LeRP algorithm yields a set of corresponding locations in the two input scene images, from which the absolute orientation may be found.

Node and edge colours are compared during the matching process. As these colours are continuous quantities, a threshold on colour differences was applied.

2.6 Determining Absolute Orientation

Horn's method [9] was used to determine absolute orientation. This reveals all 6 degree of freedom in the rigid transformation. It operated on the corresponding landmarks identified by the graph matching operation.

After an initial pass with Horn's technique, the residual error of each (3-D) landmark location is computed. The presence of outliers is checked via a simple threshold and the associated landmarks are removed and the transform recomputed. The process of finding the transform and checking consistency was repeated (fewer than) **F** times, for **F** landmarks. Hence the effort in this stage of the processing is bounded by O(**F²**). A minimum of 3 pairs of corresponding points is needed to Horn's method. Typically no

outliers were removed, but occasionally some were present and this step helped improve accuracy.

3 Summary of Computational Effort

Each of the processing steps requires worst-case effort that has a polynomial bound. The effort described in Table 2 assumes: **NxM** range images and **F** landmarks. **D** is the mean degree of landmark graphs and **Q** is the total number of edges.

Step	Sub-Step	Effort
Find Graph Vi	Predict Sensor Motion	O(Constant)
	Find Vi from Vi-1, and V0	O(F)
Find Landmarks Li	Find KLT Corner Features	O(N M)
	Refine Landmark Locations	O(F)
	Reject Unstable Landmarks	O(F)
Find Graph Gi	Find Structure Mimicking Vi	$O(QF^2)$
	Colour Nodes	O(F)
	Colour Edges	$O(F^2)$
Match Graphs Vi and Gi	LeRP algorithm	$O(F^3D^2)$
Find Coordinate Transform Ti	Use Horn's Technique	O(F)
	Remove Outliers & Recompute	$O(F^2)$
Update Graph V0	Refine Landmark Locations	O(F)

Table 2: Computational Effort for Landmark Graph-Based Registration

Processing effort for the graph matching is on the order of **$O(F^3D^2R)$,** where **F** is the number of nodes, **D** is the average degree. The parameter **R** is actually a weak function of **F** (see [4]) but was set to a constant in all tests reported herein.

4 Testing and Results

As shown in Table 3, two types of scene data were used. The 'Sensor' data was acquired using a structured light sensor, similar to [3], that was built in-house. The device includes a mechanical positioner, permitting it to collect the 3-D data set. The sensor also includes a colour camera [30].

The in-house sensor is relatively slow, relative to commercial 3-D range cameras. Furthermore, it is these state-of-the-art range cameras that are driving the goals for the new registration technique. Hence range images from the sensor were acquired and then stored for use in testing. Point clouds generated by the sensor were resampled to form scene imagery, as shown in Figure 4.

The 'Synthetic' test scenes were generated using models of blocks, with ray tracing calculations to determine each pixel of the scene. Gaussian noise was added to the range and hue images during testing, spanning 2% of the intensity range.

Tests with synthetic scenes benchmarked accuracy for both sensor translation and rotation. Translational steps were 0.1 inches – 3% of the 3.6" sensor FOV. The total shift over an image sequence was 22% of the FOV. Rotational steps were of 1 degree spanned a total range of 5 degrees. A total of 23 unique synthetic scenes were used in testing. Synthetic scenes were reused used on multiple test trials, by adding random noise. Scenes scanned with the in-house sensor had translational steps of 2.5% (of an 8" FOV) with an overall span of 30% FOV. Objects scanned with the real sensor had a

fairly consistent standoff. Hence the synthetic images appeard somewhat more cluttered.

4.1 Benchmarks Against a Fast-ICP Algorithm

An ICP algorithm was used for comparison purposes [1][18]. ICP was implemented with a simplex optimisation routine [15]. To reduce the effort of evaluating the cost function, a simple image difference was used. In each evaluation the point cloud was transformed and a range image was formed. The mean absolute difference in pixels was then computed. This avoids the step of having to search for the closest point in the point cloud. It does introduce error, however, because the closest 3-D point might actually be found via a lateral shift to an adjacent pixel. To improve compute speed only the grey level difference with the current pixel was considered. This type of resampling is also described in [24].

Another variation on ICP used herein maintained a fixed number of iterations. This permitted more direct comparisons between ICP and the landmark-graph method, as both algorithms were deterministic. Exactly 200 iterations of ICP were run in each trial.

Another speed optimisation step for ICP involved limiting the region of interest used in the cost function comparisons to areas near landmarks. This somewhat intertwined the implementation of the two techniques. This approach is being investigated as a potential means to create a hybrid technique that combines ICP with the landmark-graph method.

4.2 Test Results

Table 3 summarizes a total of 32 trials. Each 'trial' consisted of one registration operation on a given pair of images. Percent error was computed via the absolute mean displacement error, and then expressed relative to the sensor field of view. Rotational errors are also given as absolute means. All means were computed over the entire image sequence, including initial transient.

-	-	#	Length	Trans. Error		Rotation. Error		Rate: Mean +/- Std. Dev. (Hz)	
Scene	Type	Trials	Stream	Graph	ICP	Graph	ICP	Graph	ICP
Synthetic	Rotation	5	6	0.6%	0.6%	1.7^0	1.0^0	10 +/- 9%	0.13 +/- 20%
Synthetic	Trans.	16	9	0.1%	0.2%	1.7^0	0.7^0	10 +/- 6%	0.14 +/- 21%
Sensor	Trans.	11	12	0.6%	0.6%	1.1^0	0.3^0	10 +/- 6%	0.14 +/- 6%

Table 3. Results of test trials indicate that the accuracy of the landmark graph approach is comparable to Fast-ICP, while providing a substantial speed improvement. ICP was run with 200 (a constant number of) iterations. Mean absolute error values are reported. All images were 200x200; tests run on a 900MHz PC.

As seen in Table 3, the landmark graph approach rivals the accuracy of Fast-ICP, while executing much faster, at ~10Hz. The landmark-based approach also provides a greater degree of determinism than ICP as indicated by the standard deviation of the processing time. Fast-ICP was run with 200 iterations and the images for all test trials were 200x200. Tests were run on a 900MHz PC.

A reason ICP suffers variations in processing rate despite the use of a fixed number of iterations is due to the simplex-based optimisation [15]. The simplex method will use a varying number of evaluations of the cost function, depending on the type of simplex movement ('contraction' versus 'flip'). Note that other optimisation techniques also possess this type of computational variation per iteration, such as Hooke-Jeeves.

The landmark-graph approach also suffers some variations in processing rates, as shown. This is due to varying numbers of landmarks (F in Table 2). The degree of this variation can be mitigated somewhat, for example if the system processes an image stream with substantial changes to scene content, by adjusting detection threshold of KLT. Dynamic means of threshold adjustment are of interest in future studies. Limits on **F** are also possible when computing Graph Vi. See Table 2.

Figure 4. Landmarks and scene graph for range data acquired with a structured light sensor.

5 Conclusions and Future Studies

Reported results are encouraging. The landmark-based approach is able to achieve rates ~10Hz for 200x200 images on a single PC. 1% accuracy appears comparable to ICP and to other reported techniques. Yet landmark-graphs provide higher speed and greater determinism than ICP. The method also yields 6 DOF alignment parameters and can process image streams with relatively large jumps in sensor position (5%). The ability to handle large changes in sensor position not only allows for fast sensor motion, but also indicates that precise predictions of sensor motion are not critical for accurate registration.

The landmark-graph approach does have its share of challenges. Most significant is the computation of landmarks via a method that is both fast and stable. The KLT feature detector appears to work well. It would be beneficial to improve the rotational invariance. A circular neighbourhood might help in this regard. This is an on-going area of investigation.

Stable placement of graph edges was a challenge in a previous implementation [23], but the predictive approach that mimics graph structure of V0 seems much better.

Future extensions could also include marrying the new technique with ICP in a post-processing step. This would permit more scene data to enter into the final transform calculations than just the landmarks. ICP could begin iterations using the alignment parameters derived via landmark graphs. A fixed number of iterations could also be used for determinism. A goal here would be to target improved rotational accuracy and it might be best to just search over those 3 DOF. As with the implementation here, the region of interest used in cost function evaluations for Fast-ICP can be restricted using the image regions near corresponding landmarks.

A general improvement to the landmark-graph approach is planned in the near future. This involves a post-processing step, where the graph V0 is grown in size after each new image is aligned. This is an O(F) operation, see Table 2. Growing V0 in an on-line fashion would permit extended regions of surface data to be incorporated into a single contiguous data set. The implementation reported here is more suited for repeated sensor scans over a small area, from differing views.

Future work is also planned for an integrated visualization subsystem to provide a complete real-time remote viewing system.

6 Acknowledgements

References

[1] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, IEEE Trans. Pattern Analysis and Machine Intelligence, 14 (2) (1992) 239-256.
[2] B.K.P. Horn and J.G. Harris, Rigid Body Motion from Range Image Sequences, CVGIP: Image Understanding, 53 (1), pp.1-13, 1991.
[3] F. W. DePiero and M. M. Trivedi, "3-D Computer Vision Using Structured Light: Design, Calibration and Implementation Issues," AIC, vol.43 (1996).
[4] F. W. DePiero and D.W. Krout, LeRP: An algorithm using length-r paths to approximate subgraph isomorphism, Pattern Recognition Journal, 24 (2003) 33-46.
[5] O. Faugeras, 3-D Comp. Vision, A Geometric Viewpoint, MIT Press, Cambridge, MA, 1993.
[6] M. Fischler, R. Bolles, Random Consensus: a paradigm for model fitting with applications in image analysis and automated cartography, Communications of the ACM, 24 (1981) 381-395.
[7] Birch, Implementation of the Kanade-Lucas-Tomasi Tracker
http://vision.stanford.edu/~birch/klt/.
[8] B. Luo and E.R. Hancock, Structural graph matching using the EM algorithm and singular value decomposition, IEEE Trans. PAMI, 23 (10) (2001) 1106-1119.
[9] B.K.P. Horn, Closed-form solution of absolute orientation using unit quaternions, J. Optical Society of America A, 4 (4) (1987) 629-642.
[10] A. E. Johnson and S. B. Kang. "Registration and integration of textured 3-D data." Image and Vision Comput., 17, 135-147, '99.
[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.
[13] C. Kapoutsis, C. Vavoulidis, and I. Pitas, Morphological iterative closest point algorithm, IEEE Trans. On Image Processing, v. 8, no. 11, pp 1644-1646, Nov. (99).
[14] T. Tommasini, A. Fusiello, E. Trucco and V. Roberto. Making good features track better. in Proceedings IEEE Computer Society Conference on Computer Vision Pattern Recognition, 1998, pp. 145-149.
[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, NY, 1988.
[16] J.R. Shewchuk, Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator, Proc. First Workshop on Applied Computational Geometry, Philadelphia, Pennsylvania, pages 124-133, ACM (May 1996).
[17] S. Hsu, Multiple-view constrained video registration and its applications, Workshop on Video Registration, IEEE Computer Society, Vancouver Canada, (July, 2001).

[18] G. C. Sharp, S. W. Lee, D. K. Wehe, ICP registration using invariant features, IEEE PAMI, 24 (1) (2002) 90-102.

[19] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice-Hall, NJ, 2001.

[20] J. T. Kent, K. V. Mardia, J. M. West, Ridge curves and shape analysis, British Machine Vision Conference, University of Edinburgh (1996).

[21] X. Pennec, N. Ayache, J-P. Thirion, Landmark-based registration using features identified through differential geometry, Handbook of Medical Imaging, I. N. Bankman ed., 499-513, Academic Press (Sept 2000).

[22] A. E. Johnson, Surface landmark selection and matching in natural terrain, IEEE Computer Vision and Pattern Recognition, v. 2, 413-420 (2000).

[23] F. W. DePiero, "Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results," Proc. 1st Intl. Symposium on 3-D Data Processing, Visualization and Transmission (3DPVT), (Padova, Italy), June, 2002.

[24] A. E. Johnson and A. Miguel San Martin, Motion Estimation from Laser Ranging for Autonomous Comet Landing, IEEE Computer Vision and Pattern Recognition, v. 2, 413-420 (2000).

[25] J. Shi and C. Tomasi. "Good Features To Track." *Proc. Computer Vision and Pattern Recognition* (CVPR'94), pp. 593-600, 1994.

[27] E. Noirfalise, J.T. Lapresté, F. Jurie and M Dhome, Real-time Registration for Image Mosaicing, Electronic Proc. of The 13th BMVC, University of Cardiff, 2-5 September 2002.

[28] Gille Simon and Marie-Odile Berger, Real time registration of known or recovered multi-planar structures: application to AR, Electronic Proc. of The 13th BMVC, University of Cardiff, 2-5 September 2002.

[29] A. Fusiello, V. Roberto, and E. Trucco. Efficient stereo with multiple windowing. In *CVPR*, pages 858–863, 1997.

[30] F. DePiero, Structured Light Range Sensor, http://www.ee.calpoly.edu/~fdepiero.

[31] J. Hancock, et. al., Active Laser Radar for High Performance Measurements, Proc 1998 IEEE Intl. Conf on Robotics & Automation, Leuven, Belgium, May 1998.

Appendix – LeRP Algorithm for Approximating Subgraph Isomorphism [4]

     Main Routine
Input: Graph **G** with nodes g_i, **$0<=i<N_G$** and Graph **H** with nodes h_k, **$0<=k<N_H$**
Output: Mapping **m()**, that gives **h_k = m(g_i)**.
Steps:
I.   Compute powers of adjacency matrices A^R and B^R for graphs **G** and **H**
II.   **beta$_{peak}$[][] = *find_best_beta*(G,H,A^r,B^r)**
III.   Clear node-to-node mappings
IV.   For each **L, $0<=L<minimum(N_G,N_H)$**
     A.   Let **peak = 0**
     B.   For each unmapped node **g_i**
     C.   For each unmapped node **h_k**
          1.   Verify consistency of mapping **g_i** to **h_k** given current **m()**
          2.   **rho = 0**
          3.   For each mapped edge **e_{ij}**
               a)   lookup associated edge **e_{kl}** where **l=m(j)**
               b)   **beta = *compare*(i,j,k,l)**
               c)   **gamma = *compare*(j,j,l,l)**
               d)   **rho = 1 – (1-rho)(1-beta)(1-gamma)**
          4.   Next **j**
          5.   **alpha = *compare*(i,i,k,k)**
          6.   **rho = 1 – (1-rho)(1- alpha)(1- beta$_{peak}$[i][k])**
          7.   If **rho>peak** Then
               a)   **g$_{peak}$=i**
               b)   **h$_{peak}$ =k**
               c)   **peak=rho**
          8.   End If
     D.   Next **k**
     E.   Next **i**
     F.   If **peak=0** Then GoTo **END**
     G.   Let **m(g$_{peak}$)=h$_{peak}$**
V.   Next **L**
VI.   If **(L=N$_G$)** and **(L=N$_H$)** Then **G** *is ISOMORPHIC to* **H**, refer to mapping **m()**.
VII.   Else a subgraph isomorphism exists between **G** and **H**, refer to mapping **m()**.
VIII.   END

Function: find_best_beta(G,H,A^r,B^r)
     A.   For each node **g_i**
     B.   For each node **h_k**
          1.   For each edge **e_{ij}**
          2.   For each edge **e_{kl}**
               a)   **beta = *compare* (i,j,k,l)**
               b)   Save **beta$_{peak}$[i][k]=beta** if maximal for nodes i,k
          3.   Next **l**
          4.   Next **j**
     C.   Next **k**
     D.   Next **i**
     E.   Return **beta$_{peak}$[][]**

     Function: *compare(i,j,k,l)*
I.   For **$1<=r<=R$**
     A.   If **$a_{ij}^{(r)}$!= $b_{kl}^{(r)}$** Then Break
II.   Next **r**
III.   Return **$(r/N)^2$**

Surface Registration at 10Hz Based on Landmark-Graphs: Benefits for a Scalable Remote Viewing System

Abstract – *Real-time surface registration is a key technology for the development of future remote viewing systems. An architecture for a video distribution system supporting multiple users, with individual viewpoint selection, is suggested. The approach would provide a transmission bandwidth independent of the number of users, for scalability. The proposed architecture uses a method of surface registration based on landmark-graphs. Results from 141 test trials on synthetic scenes indicate that a mean absolute positioning accuracy under 1% of the sensor field of view is possible. The mean rate for registration was 10Hz, with a standard deviation under 10%. Tests were benchmarked on a 900MHz PC. The sensor images were 200x200 pixels and contained both range and color imagery.*

1. Flexible Remote Viewing Systems

The goal of this research is to further methods of surface registration, for the enhancement of remote viewing systems. Current viewing capabilities such as TV or teleconferencing are quite limited by restrictions in viewpoint as each user is fed the same view. Furthermore, the selection is restricted to discrete camera signals.

Improved remote viewing systems should provide viewpoint flexibility for multiple users. Preferably, this should be done without simply introducing a camera with pan & tilt for each user and without increasing the transmission bandwidth in proportion to the number of users.

Such improvements may be possible, given an ability to do real-time surface registration. This refers to 'stitching together' sections of a scene that have been acquired by sensor(s) from different vantage points. This permits a large contiguous set of surface data to be constructed, as a basis for rendering remote views.

Accomplishing registration in real-time means that the alignment calculations must be completed at the rate of sensor acquisition, thus permitting immediate use of the sensor data for remote viewing. Voxel-based rendering could then provide imagery with an arbitrary viewpoint.

Given the real-time registration capability new approaches to video distribution become possible. See Figure 1. The server acquires new sensor images, and then computes an alignment relative to previous inputs. By transmitting the new sensor data to clients along with alignment transformations, the rendering operations may then be off-loaded to client machines.

This approach permits each client to have an independent viewpoint. It also means that the bandwidth of the transmission is determined by the sensor(s) only, not by the number of users. The method also offloads considerable effort, by not centralizing all the processing and rendering calculations [1].

Figure 1. Architecture for a scalable remote viewing system with multiple users. View points are controlled by each user.

A visualization system should provide rapid response to user requests for new viewpoints. The proposed architecture is well optimized in this regard, as the viewpoint request and subsequent rendering are all local to the client machine. This makes the rendering frame rate and response to pan and tilt view changes all independent – and not limited by - the sensor data rate or the transmission rate.

2. Potential Applications
.
Applications such as a 'television with a joystick' would become possible, given the ability to perform real-time registration. This would support a broadcast transmission to many users, each with an independent viewpoint. For example with a sports broadcast, some viewers might choose to watch the hands of a golfer, others the ball, others the whole putting green. Scenes with an individual golfer would be amenable to this sort of remote viewing system. More complex scenes (such as a crowded street) could have a prohibitive level of occlusion, despite multiple sensors. For applications with tele-immersion, two such views could be computed, one for each eye.

Another application area is tele-medicine. A scenario is proposed here that is more flexible than just the transmission of individual medical scans. Rather, more interactive modes of observation are envisioned. For example if a field technician positioned a sensor over a patient's wound, then a remote doctor could examine the injury. Furthermore, if the doctor's viewpoint could be graphically presented to the sensor technician, then the doctor's viewing needs could be better anticipated.

In another remote-viewing scenario a robot could use the doctor's viewpoint as a basis for path planning and sensor positioning.

Awareness of another person's viewpoint is pre-attentive knowledge, when interacting directly. However, this knowledge can be lost in a remote-viewing scenario. Means to graphically present a remote user's viewpoint may be a useful feature for advanced systems.

3. Areas of Investigation

All of these advanced viewing scenarios rely on surface registration. The fundamental reason that registration is required is because sensors such as laser range finders (and even simple video cameras) are line-of-sight devices. Hence either multiple sensors or multiple images (from a moving sensor) would typically be required to form a complete set of surface data across an entire scene. Figure 2 illustrates the line-of-sight nature of a range sensor. The 2nd image has been rendered from a viewpoint that was offset from the sensor, revealing missing surface data.

Approaches for registration and visualization need to be deterministic and computationally tractable for real-time implementation. Methodologies in these areas are the focus herein. Also, this study is restricted to cases with static scenes that are scanned by a moving sensor.

Figure 2. Sensor image (above, simulated) and a scene rendered from another distinct viewpoint. This illustrates the line-of-sight nature of range sensors. A low-resolution voxel array was used to store surface data.

4. The Challenge of Surface Registration

Surface registration is the process of determining the six DOF that describe changes in sensor location between a pair of input images. The goal here is to track changes in sensor location as the device is moved continuously across some arbitrary scene. The landmark-graph approach reveals sensor motion based entirely on an analysis of scene content – using no auxiliary sensors or alignment targets.

Work in registering range data dates back to random approaches such as RANSAC [2] and iterative methods have been widely studied [3]. However, non-deterministic methods such as these are not preferred for real-time implementations. Robust methods that are computationally intensive have also been reported [4] but may not be able to achieve high frame rates. Other methods that track features [5] assume small image displacements and then use an affine motion model to describe local scene changes. The assumption of small displacements limits sensor velocity.

Some real-time methods have also been recently proposed [6][7]. However a direct comparison to these works cannot be made, as these rely on either a fixed camera position (rotation only) [6] or on an assumption of a particular type of scene content [7].

Reported methods typically do not separate the steps of determining corresponding points and determining the transform [8][3]. These steps are kept separate for landmark-graphs thanks to the LeRP algorithm for approximating subgraph isomorphism [9]. This is an important distinction with respect to computational efficiency.

5. Surface Registration and Remote Viewing With Landmark Graphs

Stability for the landmark-graph is provided via the similarity of inter-landmark geometry, which is verified via a subgraph-matching algorithm. This is in contrast to approaches such as [5] which provide robustness based on checks of deviation in the path of each individual feature, but that do not enforce a specific geometrical structure (attributed graph) between features. See Figure 3.

The result of the graph matching processing step is a pair of subgraphs with identical structure (in terms of nodes and edges). The pair of subgraphs also has attributes that match to within specified tolerances. As such, a rigid transformation may be computed between the landmark correspondences given by the matching subgraphs.

The following notation is used, to describe the processing and representation of an image stream. The stream is composed of a sequence of sensor images, indexed by i = 0, 1, 2...

- F_i, Sensor coordinate frame for i_{th} scene.
- (R_i, C_i) Range & color images acquired at F_i.
- L_i, Set of landmarks found in R_i (w/rt F_i).
- G_i, Graph formed from landmarks L_i.
- T_i, Coordinate transform relating F_i to F_0.
- V_0, Graph associated with all landmarks for entire image stream.
- V_i, Predicted subgraph of V_0, approximating G_i.

The world coordinate frame for the voxel array is aligned with F_0. Registration calculations are based on comparisons between the i_{th} sensor location, F_i, and the initial location, F_0.

In a remote viewing system based on landmark-graph registration, the server could execute the following steps:

1) Acquire new sensor image.
2) Predict V_i based on V_0 and motion estimate.
3) Find landmarks L_i in range image R_i.
4) Form G_i using L_i, mimicking structure of Vi.
5) Compute attributes for G_i, using R_i & C_i.
6) Use LeRP algorithm to match G_i to V_i, the resulting subgraph mapping gives the L_i to L_0 correspondences.
7) Find transform T_i via Horn's method, using the L_i to L_0 correspondences.
8) Compress R_i & C_i and broadcast to clients, along with T_i.
9) Update landmark positions L_0 and attributes stored in V_0. Grow V_0 using any new territory exposed in G_i.
10) Repeat

The client could execute these steps:

1) Receive R_i & C_i along with T_i. Decompress sensor imagery.
2) Accumulate R_i & C_i into voxel array using the T_i transform.
3) Repeat.

The client would also continuously render scene images, based on the current voxel array content. This could be done asynchronously; at whatever rate the client platform can manage.

Previous work by this author with landmark-graphs restricted analyses to individual pairs of sensor images, not to image streams [10]. Stream processing is more appropriate for the continuous sensor movement. With an image stream, prediction may be used, as in [5]. Results of the landmark-graph approach, including prediction, are superior to those

previously reported by this author [10]. More information on LeRP, the graph matching technique is available in [9]. As LeRP is a relatively new algorithm, pseudo-code is included in the appendix.

6. Transmission Subsystem

Given that range data is available in addition to standard intensity images, and given the alignment data, there are new opportunities for image compression for the transmitted sensor data. Because sensor data is in the form of images, some simple variation on standard image compression techniques may be useful for the remote viewing system.

For example, the coordinate transforms T_i and T_{i-1} could be used to warp the images C_{i-1} & R_{i-1} to approximate the current images C_i & R_i. An image difference operation could then provide better compression over a method such as MPEG thanks to the warping operation that would make subsequent images more similar. Note the receiver would of course have to perform an un-warping operation.

Figure 3. Landmark graph

7. Visualization Subsystem

In the proposed architecture, the client machine is tasked with accumulating range data and rendering user images. This offloads computations from the server side, making for a more balanced load. This also facilitates each user having their own viewpoint.

The method of shear-warp ray casting [11] is proposed for rendering. This method introduces a shear offset between adjacent layers effectively giving the voxel array a parallelogram shape. A projection of voxels then occurs along rows and columns of the array. This sort of projection is much more efficient than ray-tracing, for example. Projections are performed back-to-front, relative to the user viewpoint. The warp operation restores proper image aspect ratio.

The compute performance demands on client processors in this architecture may be somewhat beyond the capability of today – depending on sensor data rates and image size. However, less expensive memory, faster general-purpose processors, and voxel visualization boards [12] may all contribute to meeting these increased demands, soon.

Choices of using a voxel array, and shear-warp, were driven by the use of 3-D point clouds of sensor data and the need for real-time processing. The voxel array is well matched to the storage needs of the 3-D data points. Shear-warp then provides efficient means for rendering.

Despite the simplicity of a point cloud approach, it may have some advantages over methods that use a polygonal surface representation [13]. Consider a situation with the sensor being swept back and forth over a static scene. As new 3-D points are acquired and

aligned, a simple algorithm may be used to accumulate the data into the array – for example, just replacing the old points with new ones. Alternatively some type of averaging color values (hue) could be used when accumulating data. In contrast to this, consider a polygon-based approach. The polygons output from a sensor subsystem would have to be continuously merged to avoid unbounded growth of the surface description [13]. Such recombination and merging could be challenging in real-time. The voxel-based approach avoids this sort of problem.

8. Testing and Results

This is an on-going effort and the results of the registration with prediction are currently the main focus of investigation. Additional results documenting the effect of compression are under study.

The test suite included cases with both rotational movement and translation. Both real and synthetic sensor data has been included. Zero mean Gaussian noise was added to the synthetic sensor images.

The mean absolute position error is given as a percentage of the sensor field of view. The number of pixels across the sensor and voxel array was the same in these tests. Hence the percent error in position indicates the amount of misregistration expected in the voxel array. See Table 1.

Reports of accuracy and computational rate are given in Table 1, for both the landmark-graph approach and for a 'fast-ICP' method [10]. The fast-ICP method used a simple image difference, rather than point-by-point search for correspondence. It also ran with a fixed number of iterations (200) to yield a deterministic algorithm that is more directly comparable to the landmark-graph approach.

	Translation Synthetic Scenes	Rotation Synthetic Scenes	Translation Real Scenes
Mean Absolute Error For LG	0.1%	0.7^0	0.6%
Mean Absolute Error For ICP	0.2%	1.1^0	0.6%
LG Rate Mean +/- Std. Dev.	10 Hz +/- 9%	10 Hz +/- 7%	10 Hz +/- 6%
ICP Rate Mean +/- Std. Dev.	0.13 Hz +/- 22%	0.14 Hz +/- 19%	0.14 Hz +/- 6%

Table 1. Test results for surface registration demonstrate a faster rate and greater determinism for landmark-graphs, compared to fast-ICP.

Test results in Table 1 show relatively low errors under 1% of the sensor field of view. These mis-registration errors result in a blurring of the surface data accumulated in the voxel array. Hence these error rates of are considered good. Figure 4 shows a relatively crisp image, after the accumulation of 10 registered sensor images.

The landmark-graph method was benchmarked to be ~70x faster than ICP. Landmark-graphs also provide better determinism, see standard deviations on processing rates. These factors make the landmark-graph approach superior for a real-time system.

The processing rates are given for a 900MHz PC. Although the rates are considered good relative to other reported methods, these would still need to increase for a broadcast system. Also note that the sensor image size was only 200x200 pixels. The new method does seem promising, nonetheless, given the modest compute platform.

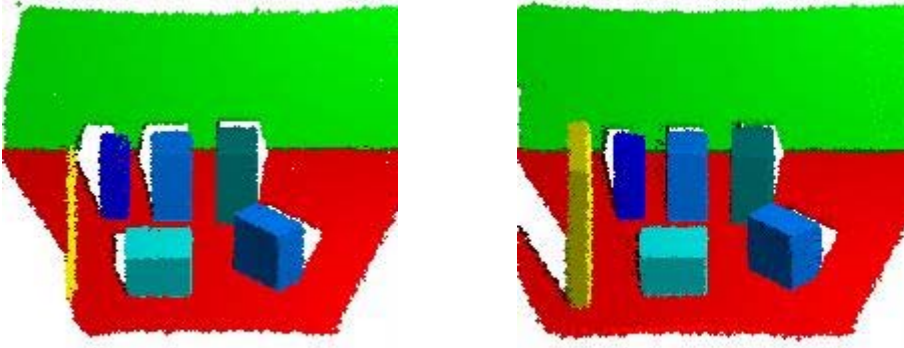AVI-format video clips are available for download [14]. The clips contain images rendered during the testing discussed below. The still image in Figure 4 is from one of these sequences.

Figure 4. Rendered image from voxel array after initial sensor image, and after the accumulation of 10 images (2nd). Note the new portions of the scene encountered after all 10 images are accumulated. Also note the reduction in the missing data (white areas).

9. Conclusions and Future Studies

Test results for the landmark-graph method of surface registration appear to yield relatively crisp imagery, with registration errors under 1%. The technique could provide the basis for a new means for distribution of surface data in a remote viewing system. Such a system could support multiple users and would be a scalable architecture. Opportunities for sensor image compression are superior to standard image streams because of the registration data, which could be used to align sensor images prior to compression.

Lossy compression methods will degrade the voxel data and the final user images, as will sensor noise and registration errors. To help mitigate some of the degradation a median operation could be performed on the voxel array. This step would retain the three most recent contributions to a voxel, and use the median of the three for rendering purposes. This and other possible post processing steps are underway.

An outstanding issue in the design of the proposed architecture has to do with the introduction of new users. If surface transmissions are underway when a new client accesses the broadcasts, then the new client's voxel array will not match the state of other clients, nor of the server. Hence some means of voxel refresh would likely be required. One possibility is to provide a secondary, non-real-time transmission from the server to the clients for this purpose. The secondary transmission might consist of only filled voxels (to reduce data rates).

10. Acknowledgements

11. References

[1] VisServer by SGI, http://www.sgi.com.

[2] M. Fischler, R. Bolles, Random Consensus: a paradigm for model fitting with applications in image analysis and automated cartography, *Communications of the ACM*, 24, pp. 381-395, 1981.

[3] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, *IEEE Trans. PAMI*, 14 (2) 239-256, 1992.

[4] A. E. Johnson and S. B. Kang. "Registration and integration of textured 3-D data." *Image and Vision Comp.* 17, 135-147, 1999.

[5] T. Tommasini, A. Fusiello, E. Trucco and V. Roberto. Making good features track better. In *Proc IEEE Computer Society Conference on Computer Vision Pattern Recognition*, pp. 145-149, 1998.

[6] E. Noirfalise, J.T. Lapresté, F. Jurie and M Dhome, Real-time Registration for Image Mosaicing, *Electronic Proc. of The 13th BMVC*, University of Cardiff, September 2002.

[7] G. Simon and M.-O. Berger, Real time registration of known or recovered multi-planar structures: application to AR, *Electronic Proc. of The 13th BMVC*, University of Cardiff, September 2002.

[8] B. Luo and E.R. Hancock, Structural graph matching using the EM algorithm and singular value decomposition, *IEEE Trans. PAMI*, 23 pp. 1106-1119, 2001.

[9] F. W. DePiero and D.W. Krout, LeRP: An algorithm using length-r paths to approximate subgraph isomorphism, *Pattern Recognition Journal*, 24, 33-46, 2003.

[10] F. W. DePiero, "Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results," *Proc. 1st Intl. Symposium on 3-D Data Processing, Visualization and Transmission* (3DPVT), (Padova, Italy), June, 2002.

[11] P. Lacroute, and M. Levoy, Fast Volume Rendering Using a Shear-Warp Factorization of the Viewing Transformation, *Proc. SIGGRAPH*. ACM, Orlando, FL, pp. 451-458. 1994.

[12] Voxel visualization board, by TeraRecon, http://www.rtviz.com.

[13] Heckbert, P., and Garland, M. Survey of polygonal surface simplification algorithms. Tech. Rep. CMU-CS-95-194, Carnegie Mellon University, 1995.

[14] DePiero, AVI Clip showing results of L-G Registration, http://www.ee.calpoly.edu/~fdepiero.

Development of an Autonomous Tactical Reconnaissance Platform

Dianne J. DeTurris
Aerospace Engineering Department

Sema E. Alptekin
Jon C. Ervin
Industrial & Manufacturing Engineering Department

Cal Poly State University

Development of an Autonomous Tactical Reconnaissance Platform

Final Report

By

Dianne J. DeTurris
Sema E. Alptekin
and
Jon C. Ervin

November 2003

1

Abstract

The Autonomous Tactical Reconnaissance Platform (ATRP) is a parafoil with a payload that is instrumented for autonomous flight and remote sensing capabilities. The purpose of this project is to prototype a system that can be used as a personal reconnaissance device. The ATRP has many military and civilian uses that are not adequately addressed by remote sensing systems currently available.

Building on work that occurred in Phase I, in this Phase II effort major advances were made in the electronics, software, launch mechanisms and the launch vehicle. The participation of undergraduate students was crucial to the success of this project and they made major contributions in all areas of design, development and test. Results from the Phase I effort and preliminary results of this Phase II follow-on were presented at the AIAA Aerodynamic Decelerator Systems Technology conference in Monterey, California.

Highlights, papers and additional information on this project including a rocket launch video can be accessed at the web site,
http://www.ime.calpoly.edu/salpteki/Parafoil/Parafoil-Index.htm.

1. Background

The Autonomous Tactical Reconnaissance Platform (ATRP) is a parafoil that carries a payload instrumented for autonomous flight while performing remote sensing activities. A parafoil is a flying wing made of a flexible material, supported and controlled by an array of strings. The ATRP has many potential military and civilian uses that are not adequately addressed by remote sensing systems currently available. Although the ATRP would provide data similar to that from either Unmanned Air Vehicles (UAVs) or Micro Air Vehicles (MAVs), it has unique capabilities and advantages that separate it from both of these types of air vehicles

Reconnaissance data from various sources have proven to be crucial to the success of most battle campaigns, particularly in recent years. Today our military employs a number of reconnaissance assets including satellites, manned aircraft, unmanned aircraft and human infiltration. All of these assets can provide vital information to commanders at home and in the field, however they all have drawbacks when it comes to supporting the individual soldier in real-time, in hostile environments.

In its military version the ATRP would provide the soldier with a close range, real-time view of the immediate surroundings. This capability would be extremely valuable in rugged, urban or jungle terrain and during nighttime operations. An infrared and/or visual sensor aboard the ATRP would pinpoint the enemy location and numbers during such exercises. However, the need for real-time reconnaissance data is not limited to the battlefield.

Non-military uses include providing fire location and egress points for forest fire fighting crews and aiding search and rescue teams in locating lost hikers and victims of avalanche. This device also shows great promise in helping farmers and land management professionals to determine areas of plant stress due to drought, pests, salt intrusion and disease. Another

2

potential use for the ATRP is to provide an enhanced capability for long-range communications in obstructed or rugged terrain. Other life saving and communications scenarios can be implemented for this device when outfitted with the appropriate instrumentation.

In the area of crop and wilderness management, a number of studies have shown the value of remote sensing data [1,2]. Current collection methods such as satellites and manned aircraft are expensive, geographically imprecise, and frequently unavailable at critical times in the growing season. The ATRP equipped with a Global Positioning System and the appropriate sensors would overcome all of these obstacles and would allow farmers to perform their own personalized data collection and evaluation.

2. Introduction

The ATRP is designed to provide personnel in the field with a means of expanding their knowledge of the near environment in a safe, timely and cost effective manner. The ATRP is launched to low altitude (typically <1000 feet) and then flies a pre-selected course while transmitting sensor data to the user on the ground. During flight, the device is fully autonomous allowing the user to direct their concentration on the sensor data (i.e. IR imaging, visual imaging, radio communication and potentially laser designation) being communicated to a hand held receiving device. Prior to launch, the user can select one of several pre-programmed flight patterns for the device to follow and then launch, all in a matter of seconds.

The parafoil design is lightweight, easily compacted and deployed, and very stable in flight. An autonomous, fuzzy logic flight controller pilots the parafoil for the ATRP. After launch and deployment, the autonomous controller executes the pre-selected program of flight maneuvers as it glides in the descent phase. A fuzzy logic control method is employed because of advantages in fault tolerance, graceful response to missing or noisy sensor input and compact memory requirements. The entire system is extremely rugged, capable of surviving hard landings and small arms fire.

The ATRP and its launch mechanism are man-portable, with set-up and launch operations achieved in a matter of seconds. A range of launch mechanisms including rocket, compressed air, artillery shell and tethered flight (flown like a kite) are possible with this device. The ATRP is intended to be an inexpensive, expendable device, suitable for a wide variety of mission scenarios. The technology used in the ATRP is commercial-off-the-shelf (COTS) components, easily acquirable and inexpensive to manufacture.

At the beginning of this Phase II effort, a parafoil flying autonomously under the direction of an electronic compass had been demonstrated. Major advances were made in the electronics, software, launch mechanism and launch vehicle. Design, simulation and bench testing were performed on an improved electronics package that incorporated an inexpensive compass and a micro electro-mechanical system (MEMS) rate gyro. This new electronics package was assembled into a separate package and is referred to in this report as "Version 2". A trade study of the "Version 1", essentially a holdover from the Phase I effort, and Version 2 was performed with results reported in Section 4 of this report. A miniature remote camera

system was also bench tested and proven to be adequate for demonstrating real-time visual telemetry.

Software was developed for the ATRP microprocessor to include inputs from all of these additional electronic sensors. In parallel with the physical hardware efforts, a simulation model that included all of the new sensors was built and run in the MATLAB/SIMULINK software suite. In addition, bench tests were performed on a Global Positioning System (GPS) unit and theoretical inputs from this device were also included in the simulation. This simulation was used extensively to model and optimize the performance of the physical hardware.

The Compressed Air Launch Mechanism (CALM) was developed to provide an alternative to rocket powered launches. The CALM was used to launch two different, half scale vehicle designs, one of which attained a measured altitude of 751 feet. This altitude is more than sufficient for testing purposes and was achieved well below the full power potential of the CALM. In the most recent test, a full-scale design was flown with a parafoil recovery system on board. In future testing, the CALM will be an invaluable tool for performing rapid turn around launches here on the Cal Poly campus.

The control software was rewritten for this Phase II effort in order to incorporate an advanced fuzzy logic algorithm known as Combs' method and to create a hierarchical structure. Using Combs' method helps to mitigate the problem known as exponential rule explosion that has plagued previous efforts in fuzzy algorithm design. As new sensors were added to the system it was found to be more efficient to produce a hierarchical fuzzy control mechanism. Both of these software advances were modeled and optimized in the simulation environment.

Flight tests were performed both at Camp Roberts, north of Paso Robles, California, and at the Dairy-Aire rocket launch event near Fresno, California. Design improvements were made to these rockets, but progress was hampered by the limited number of flight windows available to the project. This is a problem has now been alleviated by use of the CALM.

Details on the development of the rocket launch mechanism are provided in Section 3. Design and development of electronic hardware is summarized in Section 4. An overview of the development of the Fuzzy Logic Algorithm and its optimization through simulation is provided in Sections 5 and 6. Design and development of the alternative launch mechanism is presented in Section 7. Results of flight tests can be found in Section 8, followed by conclusions and recommendations for future work in Section 9.

3. Rocket Development and Construction

In Phase I of this effort, a rocket launch system was developed for flight-testing of the ATRP. One of the design goals for an operational, man-portable system was for the rocket to be less than 3 feet in length. Alternative launch mechanisms (e.g. compressed air launch, artillery shell, tethered launch and deploy) may require that the system be even more compact than the current rocket launched design. However, a rocket length of 3 feet or less should prove to be easily manageable for one person to carry even in rugged terrain. This constraint in

4

length also helped to determine various other constraints such as; system weight, rocket motor size, parafoil sizing, and maximum altitude of flight and deploy. It was decided to relax the 3-foot constraint in the development phase. Rocket designs became more on the order of 4 feet in length, making it easier to access the various components within the rocket body.

A standard motor burnout delayed ejection charge was used to separate the nose from the rocket body and release the parafoil. In order to control costs, only commercially available solid rocket motors were purchased and used in test flights. These motors were sized to allow the rocket reach to apogee and deploy the parafoil at an altitude of between 500 and 2000 ft., depending on the goals of each particular test flight.

The parafoil used exclusively in Phase II is marketed as a radio controlled powered parafoil. This new parafoil with the gas engine is shown flying in Figure 1. The product consists of a small radio controlled, gas powered engine that is suspended below a parafoil canopy. The parafoil proved to be far superior to earlier models in several important respects. The new parafoil was designed for free flight and therefore had the correct line lengths to give the proper angle of attack while flying. This model also had fewer lines attached to the parafoil, thus reducing the possibility of tangles. It also was designed for the control lines to angle much closer together than the previous parafoil used, which reduced the length of the separator rod needed for stable flight.

Figure 1 - The New Parafoil as delivered with Suspended Motor

The much smaller separation distance required for the control lines made it possible to launch the rocket with the separator rod already deployed. The rocket with the separator rod in position is pictured below in Figure 2. The added drag with the rod in this position was compensated for by the motor and thus eliminated previous problems encountered in rod deployment. Another improvement made from early designs was the significant reduction in weight achieved by using lighter materials and better construction techniques.

4. Electronic Hardware Configurations

One of the goals of the Phase II program was to explore alternative hardware configurations in order to optimize the balance between functionality and system cost. In this effort two electronics designs evolved as candidates for ground and flight test evaluations. Each system is discussed in detail in the sub-sections below. These separate configurations are referred to simply as Version 1 and Version 2. The camera surveillance system is common to both Versions 1 and 2 and is discussed separately in another sub-section.

Version 1 System

This baseline configuration has been carried over from the Phase I effort with some improvements in electronic components and software. The system consists of the Handy Board as controller, a relatively expensive electronic compass for guidance and an electronic accelerometer that provides data for structural analysis purposes [3]. This system was described in the Phase I proposal and that description is repeated in the following paragraphs.

The control function is provided by a Motorola M68HC11 microcontroller on a Handy Board with a system clock speed of 2 MHz. The Handy Board also has 32 kilobytes of battery backed RAM memory for user programming. For this project we also purchased the optional extension, the Expansion Board. The Handy Board/Expansion Board is equipped with one RS-232 serial port, 21 analog input ports, 8 digital input ports, 9 digital output ports and 6 servo motor output ports. Programming is performed on a PC and the resulting code is downloaded over the RS-232 port via a separate Interface/Charger board (Figure 3) to the Handy Board micro controller.

Figure 2 - Improved Parafoil Control Design (note the small control line separator mounted sideways through the rocket)

A Precision Navigation Inc. TCM2-50 digital compass was used to provide directional input to the Handy Board (Figure 3). An RS-232 communications channel was also needed for the Handy Board to communicate with this compass. Fortunately the Handy Board manufacturers provide a small amount of board space to accommodate additional, user-developed circuitry. A circuit design incorporating a Maxim MAX232CPE interface chip and associated electronics was developed and soldered into place on the board by project

6

team members. This circuitry was configured to allow the Handy Board to communicate with the compass. Snippets of software code, that control the acceptance of input data from the compass, were obtained as freeware from various sources on the internet [4-8]. These pieces were combined and modified by team members to perform the necessary communications task.
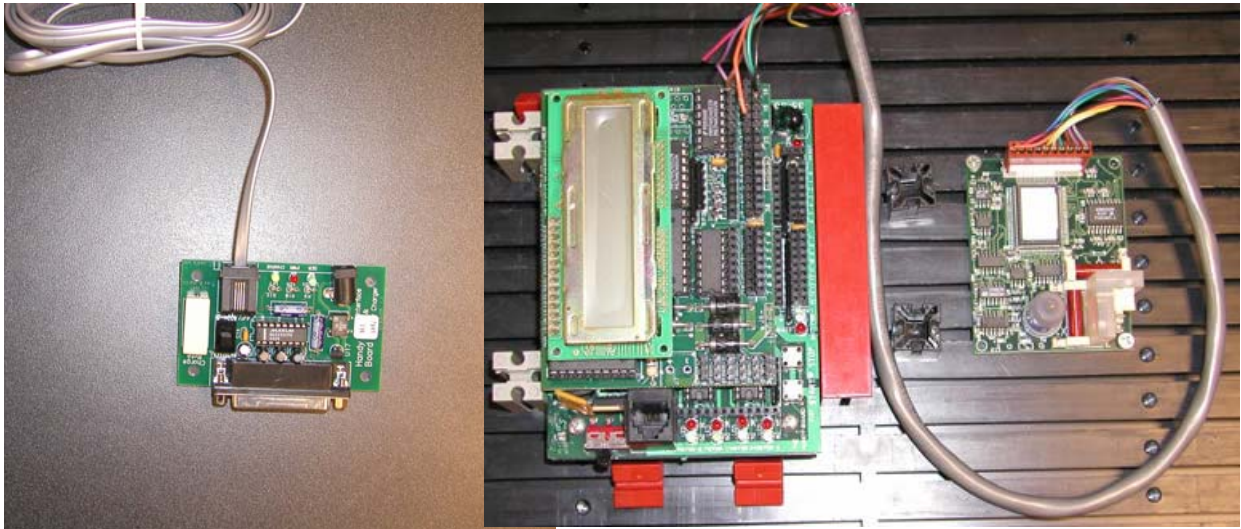
Figure 3 - Interface/Charger Board (left), Handy Board and TCM2-50 Compass (right)

The TCM2-50 compass was the single most expensive electronics component purchased for this project. The TCM2-50 spec sheet claims an accuracy of ±1.5° RMS even when tilted from level by as much as ±50°. The compass is capable of providing both digital and analog output, but is considered to be more accurate in the digital mode. It also has several auxiliary outputs including; temperature, pitch and roll angle, and three axis measurements of the local magnetic field.

Once the compass has been positioned in the electronics assembly package, it must be calibrated to eliminate local magnetic distortions. These distortions arise from close proximity to ferrous materials and magnetic fields surrounding various electronic components and support structures. In calibration mode, the compass is tilted and rotated through one or more full circles thus allowing the compass to collect magnetic field data in three dimensions. The three dimensional map of the distorting fields is computed and stored in the compass electronics.

An Analog Devices, Inc. ADXL190 accelerometer was also included in the electronics package (Figure 4). This device is capable of measuring forces up to ±100 g's along a single axis. Although output from the accelerometer was not used directly as input to the guidance package, data collected from this instrument was stored on the Handy Board to aid in future design efforts. Early in the ATRP design, the accelerometer was considered as a means of establishing when the deployment charge had been fired. This concept was discarded and a simple magnetic switch located at the separation seam was used to signal that the deploy charge had fired.

7
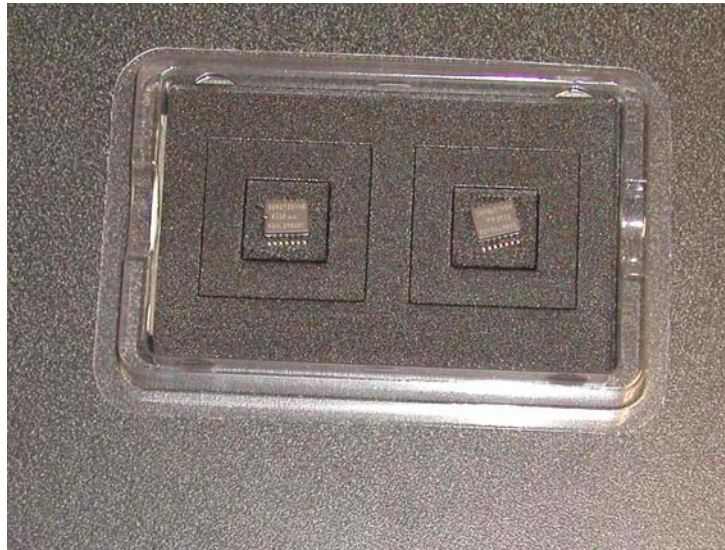
Figure 4 - A pair of ADXL190 accelerometers provided courtesy of Analog Devices, Inc.

Communication checks between the various electronic components were carried out on the bench top. Figure 5 shows all of the various flight components connected for testing and software download. Simulations of the hardware, performed in the MATLAB/Simulink software environment [9] were verified by simulated flights performed on the breadboard configuration pictured below. This synergy between computer simulation and breadboard verification played a vital role in the overall optimization process.
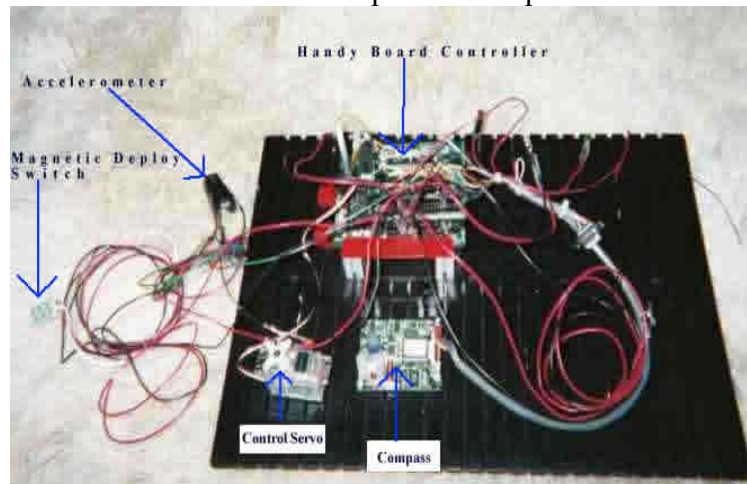
Figure 5 - Electronic components connected for breadboard testing

Assembling the various components into a rugged package that would fit within the confines of the rocket body was a subject of significant development effort. The eventual package design consisted of several lightweight bulkheads joined together by 4 threaded metal rods. The original steel joining rods had to be replaced with brass when it was determined that the steel interfered with the magnetic compass calibration. Figure 6 shows the electronics package prior to insertion in the rocket.

8

One advantage to this electronics package version is that it required only one instrument, the compass, to provide the input necessary to steer a course for the ATRP. This compass is rated to have better accuracy than the cheap compass of version 2 and therefore could be expected to provide a more accurate heading input to the fuzzy logic software. This compass can also compensate to some degree for nearby interfering metal structures. This feature would allow for flexibility in the physical layout of components.

The major drawback of this version 1 is that the TCM2-50 compass is three times the cost of the rest of the system components combined. Another potential drawback was expressed by Dr. Juan Cruz at the AIAA decelerator conference in Monterrey, CA after the presentation of our paper [10]. He pointed out that the TCM2-50 has a fluid-filled sensor used by the compass to correct for measurements made at various tilt angles. This tilt sensor works well in relatively static environments, but may induce error when subjected to moderate to high centripetal forces. Such high centripetal forces might be expected during sharp turns and buffeting experienced by the ATRP. We believe that this false tilt reading is not a significant source of heading error, due to the robust nature of the fuzzy algorithm and compensating effects in the vehicle dynamics. However, it is an effect that would have to be investigated further if the version 1 design was considered the winning candidate for further development.
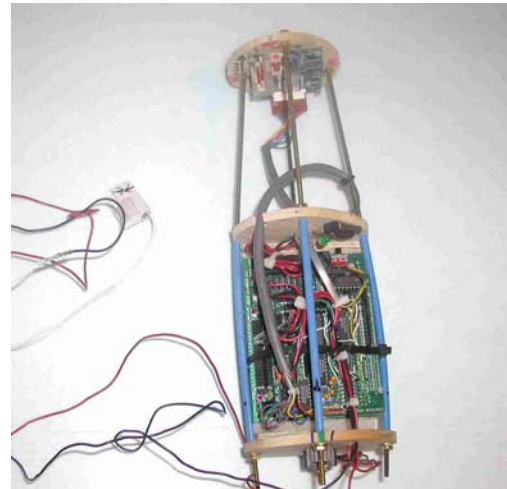
Figure 6 - The electronics package prior to insertion into the rocket

Version 2 System

This system again uses the Handy Board, as described above, to provide the software control function for the ATRP. However, it is completely different in the rest of the electronics suite and in the control software from that of Version 1. Version 2 uses a much simpler and much less costly compass, the Dinsmore Analog Sensor 1655, as well as a rate gyro to provide accurate turning velocities.

At the end of the Phase I effort one of the major unknowns remaining was how important the accuracy of the compass would be to system performance. A major factor affecting compass accuracy would be the amount of tilt from horizontal the unit experienced during flight. However, video of subsequent parafoil flight tests suggested that these tilt angles were not as severe as expected. In Phase II of this project, the feasibility of replacing the TCM2-50 (which was used in Phase I) with a much cheaper compass was investigated. The Dinsmore 1655 has only limited tilt compensation capability and is rated to be less accurate than the TCM2-50. However, the Dinsmore compass comes at a cost of $38 versus the $750 price tag for the TCM2-50. In addition to a cost advantage, the Dinsmore model is very compact and rugged.

The Dinsmore Analog Sensor 1655 is based on Hall-effect technology to detect the horizontal flux pattern of the earth's magnetic field. The output from this device is in the form of two voltages that vary roughly between 1.9 and 3.1 volts depending on the physical orientation of the device. The two voltage outputs each describe a sine like curve separated by a 90-degree phase shift as the device is rotated through a full 360-degree turn. In order to obtain an unambiguous compass heading, the voltage outputs must first be scaled and input to arc sine and arc cosine functions. The arc sine and arc cosine results are then processed to determine the current magnetic compass heading.

The two voltage measurements from the Dinsmore compass are read as analog inputs by the Handy Board where values from 0 to 5 volts are assigned to bin numbers from 0 to 255. In Figure 7 it is shown that the limited dynamic range of the raw compass voltage outputs (~1.2 volts) results in a relatively flat response over the angle range of 0 to 360 degrees. This flat response limits the accuracy of a given compass measurement. In addition, measurement accuracy is hampered by the fact that the raw compass outputs are neither exactly 90 degrees out of phase nor do they describe perfect sine functions.

The dynamic range problem was solved by Tim Reed, a student in the Electrical Engineering department, with a separate circuit that he designed and fabricated. This circuit provided individual variable controls to amplify and center each output to the full 0 to 5 volt range for the Handy Board analog inputs. A special test fixture was then developed to measure the compass response in 15-degree increments for a full 360 degrees. Theoretical sine and cosine curves were then adjusted in amplitude, X coordinate shift, Y coordinate shift and stretched/shrunk to achieve the best possible fit. The results of this fitting process for one of the six compasses tested are shown in Figures 7 and 8 without and with amplification, respectively.
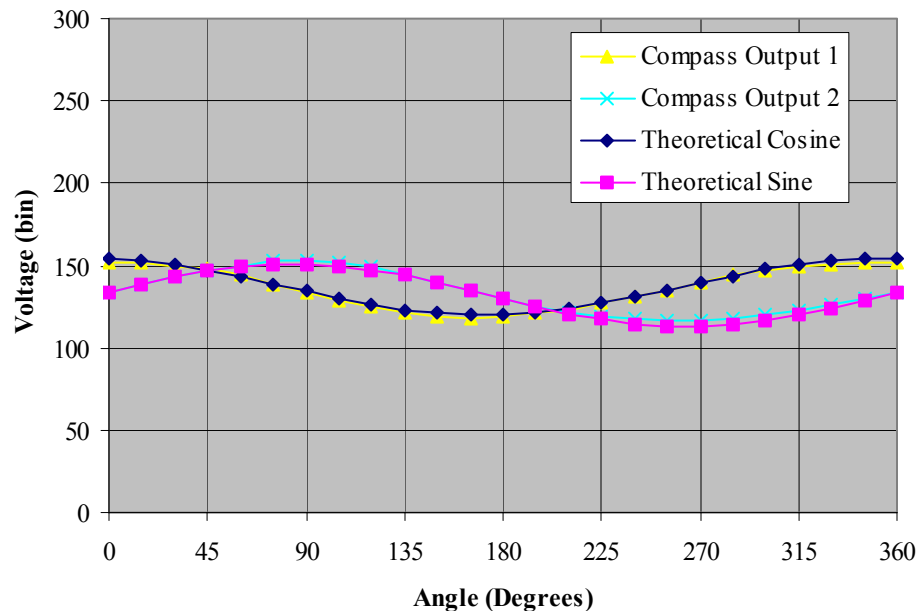
Figure 7 - Actual vs. Curve Fit Compass Outputs without Amplification

10
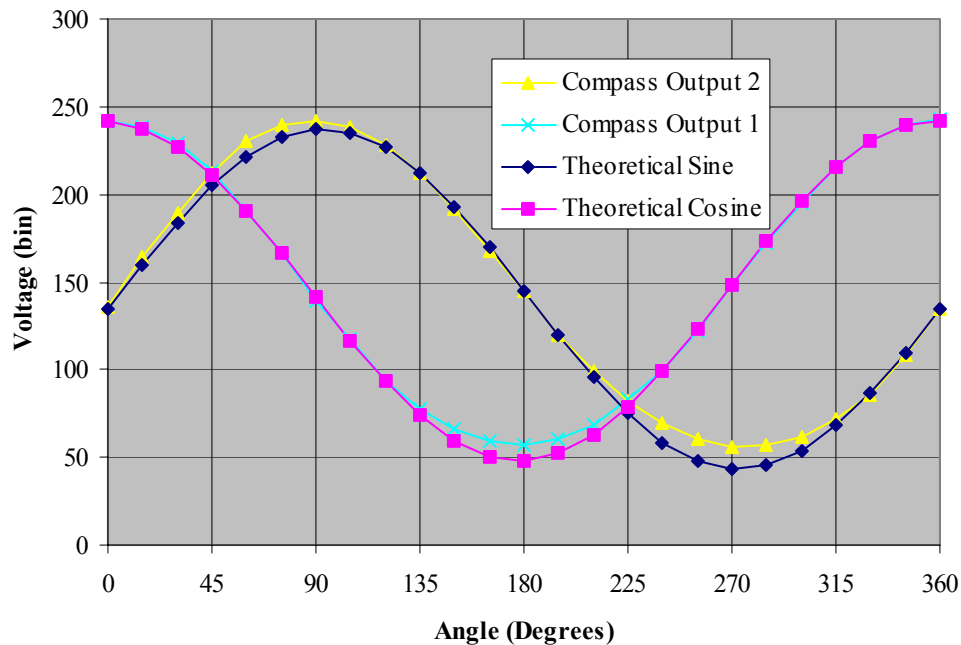
Figure 8 - Actual vs. Curve Fit Compass Outputs with Amplification

The theoretical curves are modeled in the onboard software to allow the ATRP to calculate a heading given any arbitrary compass output. Note from the figures above that the greatest discrepancy between the actual and theoretical curves occurs at peaks and valleys. The ATRP software does not use the data from these peaks and valleys, but rather selects the straight portion of one curve or the other to determine the compass heading.

The Analog Devices ADXRS150 rate gyro is a micro electro-mechanical system (MEMS) device that, when combined with the Dinsmore compass, replaces the function of the much more expensive TCM-2 compass of Version 1. The ADXRS150 is capable of measuring rotation rates of up to 150 degrees per second. It is oriented within the ATRP so rotation about the yaw axis can be measured and used as input to the fuzzy logic control algorithm. The rate gyro cost was ~$60, making the combination of the cheap compass and gyro less than one seventh the cost of the expensive compass of Version 1.

In both bench top and MATLAB/Simulink testing the two electronics versions performed with a similar degree of accuracy. It was discovered early on that two compasses could not be located any closer than about six inches of each other due to mutual interference. Originally it was thought that up to five compasses would be used to add accuracy and redundancy to the system, however the interference effect makes it difficult to have more than two compasses onboard the ATRP. Fortunately, the circuit developed by Tim Reed improved the accuracy of the cheap compass enough to obviate the need for more than one compass.

A Garmin Global Positioning System (GPS) model GPS 16 LVS was purchased to provide accurate position data for the ATRP. This GPS unit can track and use up to 12 satellites in

11

the GPS satellite system to provide an absolute position error of less than 15 meters 95% of the time. The GPS was first bench tested with a direct connection to the PC and was found to meet or exceed all stated specifications.

A bench top circuit was also developed to interface the GPS unit to the Handy Board Microcontroller. Software was written to read the input from the circuit, allowing for direct communication to the Handy Board. The readings from the GPS unit in this configuration were again checked for accuracy and found to be well within stated specifications.

Camera System

A miniature camera system was purchased to show that a relatively inexpensive video sensor could provide useful reconnaissance data. Video data from the camera is transmitted to a hand-held ground unit for real time observation and recording purposes. In an operational scenario, the ground station would be no bigger than a tablet PC, including display and receiver electronics.

This system, shown in Figure 9, features; a full color CCD camera with a built-in audio microphone and antenna, a receiver with attached patch antenna, cables for attaching to a display device and a charge converter for the receiver. The electronics that would ride on the ATRP, the camera, transmitter and transmit antenna, are incredibly light having a combined weight of only 1.5 ounces. These electronic components all run on a 5 volt battery, compatible with other ATRP electronics, and has a transmit power of 200 milliwatts. The camera transmits at the 2.4 GHz and has a range more than sufficient for the purposes of this project. The receiver and antenna are also extremely lightweight and are ideal for use in remote locations. Ground tests with this unit have shown excellent performance.

Figure 9 - The Camera, Receiver and Hardware

This unit has not yet flown on the ATRP due to a backlog of flight test priorities. The development of the CALM is expected to enable more flight tests to happen quickly. One aspect of the camera system that has not been tested is how well it will stand up to the high g forces of a launch from the CALM. However, the unit is both light and rugged and is expected to have no difficulty in this regard.

12

5. Comb's Method and the Fuzzy Logic Algorithm

A fuzzy logic control algorithm provides the decision-making strategy for the ATRP. There are a number of other excellent control methodologies available and indeed proportional integral derivative (PID) algorithms of one type or another are used in 90% of current controllers. PID controllers have been enormously successful in a variety of applications and in most cases have proven to be both very efficient and accurate. However, the fuzzy logic controller does have distinct advantages for our particular application. These advantages include robust performance in the presence of noise and signal errors, ease of development, and the capability to provide nonlinear output when required.

One of the disadvantages of a fuzzy logic solution is what is known as "exponential rule explosion". In a fuzzy algorithm, each input parameter has an associated fuzzy membership set as shown in Figure 10. In the typical fuzzy algorithm rules are formed by combining every element in the input membership sets in all possible combinations with the membership set(s) of the output variable(s). An example of rules generated from the input variables "compass heading" and "heading rate of change" and the output variable "heading correction" would be:

>   *If compass heading is "minus big" and heading rate of change is "zero"* **then** *heading correction is "plus medium"*

>   *If compass heading is "zero" and heading rate of change is "zero"* **then** *heading correction is "zero"*

>   *If compass heading is "plus big" and heading rate of change is "zero"* **then** *heading correction is "minus medium"*

>   •
>   •
>   •
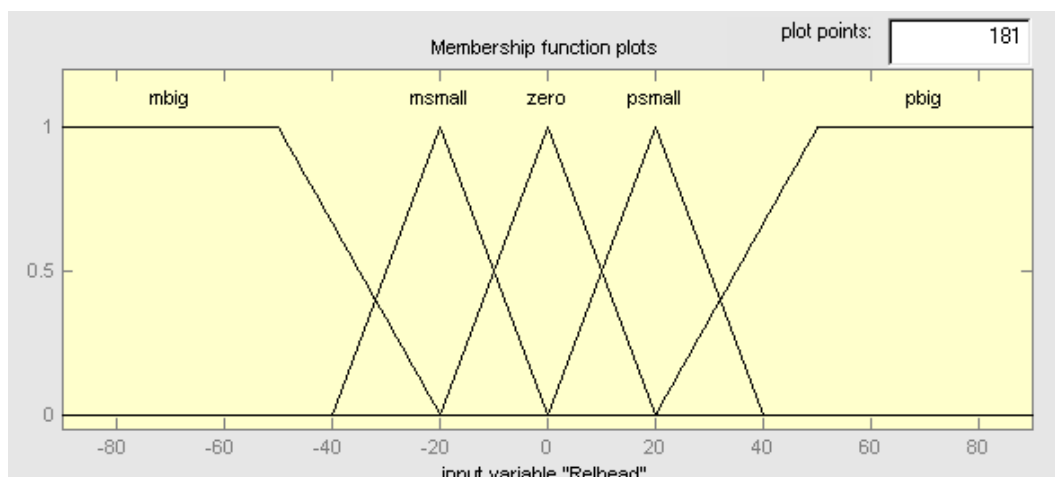
Figure 10 - Membership Set for Compass Heading Input

This system of rules works well for small two input problems, but as the number of inputs and corresponding membership sets increase linearly the number of rules increase

13

exponentially. Combs' method is a relatively obscure technique to overcome this exponential rule growth problem. For those interested in the details of this method, Cox provides an excellent introduction [11]. The advantage to this technique is that rule growth increases only linearly with the increase in the number of inputs. Therefore a fuzzy system of 5 inputs with 7 membership functions each will have only 35 rules (5x7=35) rather than the 16,807 (7^5) rules of the conventional technique. Rules formulated in the fashion of the Combs' methodology are constructed as follows:

If compass heading is "minus big" **then** *heading correction is "plus big"*

If heading rate of change is "zero" **then** *heading correction is "zero"*

If compass heading is "plus big" **then** *heading correction is "minus big"*

- •
- •
- •

The main difference between this formulation and that of the traditional method is that in Combs' method every rule has only one antecedent (the part before the **then**) for every consequent (the part after the **then**). At first it may seem counter-intuitive that this simple change results in a dramatic reduction of rules. To illustrate the savings, suppose you have two inputs of five membership functions each (membership functions for the input "compass heading" would be; minus big, minus medium, zero, plus medium, plus big). In order to cover all possible contingencies in the traditional method, every membership function for "compass heading" must be combined with every other membership function in "heading rate of change" to form 5x5=25 rules. In Combs' method every membership function of each input has a single rule to relate it to the output, resulting in 2x5=10 rules.

The breakthrough that Combs has made is to prove that his simpler formulation is equivalent to the traditional method and covers the same solution space as that of the traditional method. This incredible achievement has certain theoretical restrictions that Combs and his collaborators have outlined in a series of published papers [12-15]. In the tradition of Cal Poly's applied "learn by doing" approach, we have taken certain liberties with these restrictions in our own formulation.

Our approach is not theoretically complete in that it is not strictly equivalent to the traditional method. By performing large numbers of simulation runs, exhausting the input space, we have found no anomalous behavior exhibited using our modified formulation. These simulations have given us confidence that our current approach is sound, however because we are violating the theoretical restrictions, additional testing must always be performed any time the membership functions for the inputs are changed.

Another technique used to reduce the number of rules was to implement a hierarchical fuzzy algorithm. The inclusion of the GPS system allows the ATRP to adjust for absolute position errors introduced by winds, and asymmetric vehicle dynamics. A separate fuzzy logic algorithm uses the current and prior GPS locations to determine the effect of these external

14

forces to generate a heading compensation factor. This compensation factor is then input to the second fuzzy algorithm along with the absolute compass heading and heading rate of change to output a steering command. These two fuzzy algorithms could be combined into one, however the combined algorithm would be larger than the sum of the two existing ones.

There is very little precedent in the literature for using the Combs' reformulation of the fuzzy algorithm in control problems, however we found it worked quite well for our application. Indeed, simulation runs showed excellent performance regardless of which rule formulation was used to test the system. The memory and speed savings of the Combs' formulation were a key breakthrough for the success of the ATRP project.

6. Simulation

The parafoil design is inherently a very simple and stable flight vehicle that does not require a vertical control surface. Turns are initiated by pulling on one of two control lines connected to either side of the parafoil wing. One simplifying assumption made in the simulations was that the parafoil flies at a constant speed with no need for pitch axis control. This assumption is at least valid to a first degree of approximation and was not a limiting factor in optimizing the control algorithm. Figure 11 shows a high-level block diagram of the simulation model.
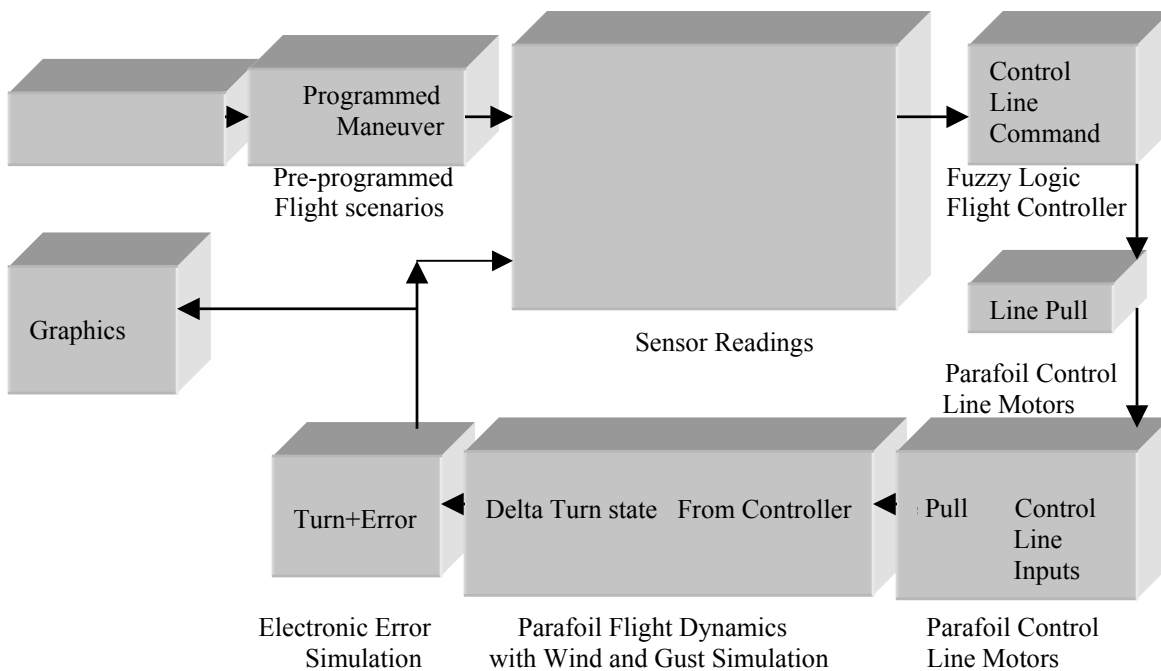
Figure 11 - Block Diagram of Simulation Model

The Matlab/Simulink software environment has been used to run hundreds of simulations of the ATRP system. The main purpose of the simulation runs is to determine the optimum architecture for the fuzzy logic control algorithm. A number of different elements, within the fuzzy algorithm and external environmental conditions, were varied during this optimization process.

Elements subject to optimization within the fuzzy algorithm included; the number and type of inputs, the number, type and value of input membership functions, the number and composition of rules, the number, type and value for the output membership functions, and the defuzzification method employed. Factors that were external to the fuzzy algorithm were modeled to determine how robust the control mechanism would be under non-ideal conditions. These external factors included; wind speed, direction and variability, system noise, instrument errors, initial deploy delay interval, and initial deploy direction of travel.

A number of major improvements have been made to the simulation model since it was first created during the Phase I effort, but the most significant change has been the addition of the GPS modeling capability. The goal of the Phase I effort was to produce an algorithm which would allow the ATRP to follow a pre-programmed course based on compass headings. In Phase II we have developed the capability to follow a course based on pre-programmed, earth coordinate based waypoints. The difference can best be illustrated when wind is introduced into the simulation model as shown in Figures 12 and 13.
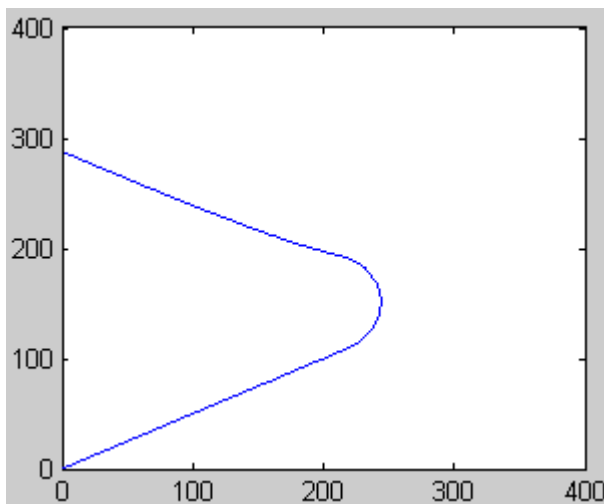
Figure 12 - Ground Track w/o GPS and 5 m/s Cross Wind
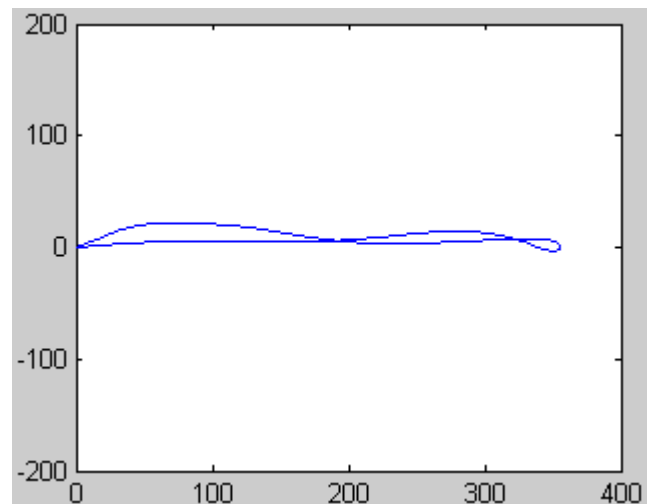
Figure 13 - Ground Track with GPS and 5 m/s Cross Wind

In both figures above, the simulated ATRP starts at the point (0,0) and is programmed to fly due West (along the positive X axis) then turn and head due East (along the negative X axis). Also in both cases, a simulated crosswind blows due North (along the positive Y axis). The difference between the two figures is that a heading correction from the GPS device has been incorporated into the simulation in Figure 13.

Without a GPS correction, the ATRP orients itself to head due West (or East after the turn is executed), but the crosswind carries it toward the North. The GPS unit provides a heading correction once every second to compensate for the wind drift. This feature allows the user to specify specific ground coordinates for the ATRP to cover while it automatically compensates for any wind condition.

16

7. Compressed Air Launch Mechanism

An alternative launch mechanism for the ATRP was designed, built and tested during Phase II. The original impetus behind this effort was to demonstrate the ability to replace the need for rocket motors in situations where they would be uneconomical, dangerous or unavailable. A trade study was performed and the Compressed Air Launch Mechanism (CALM) was determined to be an attractive alternative. In the case of farmers, compressed air would normally be readily available, could be continuously renewed and would be cheaper than buying multiple rocket motors.

As the project progressed, it became clear that an alternative launch mechanism would also be desirable for accelerating the pace of test flights. Although the PI was successful in negotiating with Camp Roberts to perform test flights there, this still did not equal the convenience of testing on the Cal Poly campus. However, we were constrained by safety considerations from using the large motors necessary to obtain usable altitudes on the Cal Poly campus.

Industrial Engineering student Andre Von Muhlen spearheaded the design and development of the CALM device, shown in Figure 14. Two tanks made of PVC pipe hold the compressed air that powers the CALM. A high-speed valve releases air from the tanks into the barrel of the device, propelling the payload to the desired altitude. Testing performed with the CALM pressurized to 1/3 capacity achieved a payload altitude of 751 feet.

Two separate half-scale projectile designs were tested with the CALM device. Aerospace Engineering student Dan Macy, shown in Figure 15, was responsible for the design and construction of both test projectiles. Each projectile had an on-board altimeter to record the maximum altitude achieved and a simple parachute for recovery. These prototype vehicles were designed to test aerodynamics of the projectile concept and to test the launch power of the CALM.

In Figure 16 Dan Macy and Dr. DeTurris are shown packing the parachute for the first projectile design to fly from the CALM. This design has retractable fins that deploy once it is clear of the barrel of the CALM. In order to get maximum lift from the CALM it is necessary to have a close fit of the body within the barrel. The retractable fins worked well on this model, however the projectile began tumbling well before achieving the maximum theoretical altitude. The cause of the tumbling was quickly diagnosed as having the projectile center of gravity to far back from the nose of the vehicle.

The second projectile, as pictured with Dan Macy in Figure 15, was a design resembling a fat arrow with no moving parts. Having learned from the first launch, weight was added to the nose of this projectile prior to launch from the CALM. This projectile had a perfect launch and exceeded expectations by flying to a maximum altitude of 751 feet. Within the 10-foot length of the barrel of the CALM, the projectile survived a maximum theoretical acceleration of 80 times the force of gravity (80 g).

A subsequent test was performed with a full-scale projectile based on the design shown in Figure 14. This vehicle was equipped with a parafoil recovery system with deployable arms

17

to separate the support and control strings. Launch and parafoil deploy were successful with this vehicle, however the deploy charge proved to be too energetic resulting in damage to one set of support strings. Future testing will employ a smaller charge as well as other lessons learned during this first test launch.

Figure 14 - Student Andre Von Muhlen and the CALM

Figure 15 - Student Dan Macy and CALM Projectile

Figure 16 - Dr. DeTurris and Dan Macy with CALM Projectile

18

These tests have proven the CALM concept as an alternative launch mechanism for the ATRP. In the event that further funding is provided for this project, the CALM will be an invaluable tool for performing large numbers of test flights here on the Cal Poly campus. Future flights would incorporate a fully operational ATRP design.

8. Test Flights

Flight tests were performed at locations outside of Fresno, CA and at Camp Roberts, CA. Camp Roberts personnel were especially helpful in making their facility available for this project. This location was much more convenient and we owe a special thanks to Major Larry Sanders for his efforts on our behalf.

The first launch, shown in Figures 17 and 18, was conducted at Camp Roberts, which resulted in a particularly satisfying success. In this test the rocket performed flawlessly in the launch and ascent phase. The deploy charge fired somewhat later than optimum causing the parafoil to experience rather high initial aerodynamic stress. The otherwise successful parafoil deploy event was marred by a folding of the parafoil wing on the right side (Figure 18) that persisted throughout the flight. However in spite of this less than ideal flight configuration, the fuzzy logic controller was able to steer the ATRP through its pre-programmed course.

Figure 17 - Launch from Camp Roberts

Figure 18 - Parafoil Under Autonomous Control

The fact that the controller was able to continue the mission under no-ideal flight conditions gave satisfying validation of the robust nature of the fuzzy logic algorithm. This success also helped to validate the computer simulation models performed throughout this effort. The delayed ejection charge issue was solved in subsequent test flights through the use of an apogee sensing system rather than relying on a time delay system.

Subsequent test flights performed both outside of Fresno and at Camp Roberts were uniformly flawless in both launch and ascent phases. As mentioned, the early problems with the deploy timing were also solved in these subsequent flights. However a new problem

arose as these other issues were solved. The parafoil often suffered from tangled lines in subsequent flight tests, making it impossible to control the descent. This issue was particularly frustrating in light of earlier flights where the parafoil flew free of any tangles. This condition persisted up to the last rocket powered test flight and will be addressed in any subsequent test project phase.

It is expected that in the future this problem will be eliminated with advances in parafoil design and development efforts. The use of the CALM in future testing will allow for much more frequent testing of parafoil designs.

9. Conclusions and Future Work

Excellent progress was made during this Phase II effort in the development of a fully autonomous parafoil platform. The PI's published a paper through the American Institute of Aeronautics and Astronautics (AIAA) detailing preliminary results obtained during this effort [10]. A subsequent presentation given to the 17th annual AIAA sponsored Aerodynamic Decelerator Systems Technology Conference in Monterey, California was well received. Feedback from attendees at this conference was helpful in guiding the further development of the project.

New rockets designs were built and tested in launches performed at Camp Roberts and Fresno. Most of the work in designing and building these rockets was performed by students in the Aerospace Engineering department of Cal Poly. Their involvement proved to be invaluable to the project and was also an excellent hands-on learning experience for the students.

Achievements in both software design and model simulation were quantum leaps over the accomplishments made in the Phase I effort. A new fuzzy algorithm based on Combs' method was developed to replace the previous fuzzy architecture. The Combs' method was modified to suit the needs of this project and subsequent simulations proved out the validity of the changes. The MATLAB/Simulink simulation was also completely revised to include input from a GPS sensor. Input from the GPS was sent to an additional fuzzy logic block to test an efficient hierarchical system for control of the ATRP.

An alternative launch mechanism, the CALM, was also designed and built during this phase. This device was tested on two half-scale and one full-scale projectile with excellent results. This device is a prototype of a system that would be employed in situations where rocket launches would be dangerous, expensive and/or inefficient. The CALM will also be the launch method of choice during future development efforts.

Major additions were made to the electronics package that provides the brains for the ATRP. Two separate electronics packages were designed, built and tested with much of the work being performed by Cal Poly undergraduate students. The design made from less expensive components was determined to be the clear winner from analysis performed on the bench top and in computer simulations.

A conference call with Dick Benney of the Natick Soldier Center, Army was most valuable in helping to guide us in future research options. Mr. Benney attended the ADS conference in Monterey and has asked to be kept up to date on our efforts. He has also provided volumes of data from other government funded efforts and we appreciate his support. Helpful suggestions were also provided by ONR representatives Fred Belen and Cliff Anderson during their visit to Cal Poly on September 30, 2003 and their ideas will be incorporated into the proposal for the next phase of this project.

Future work will include the total redesign of the microcontroller electronics hardware. The Handy Board microcontroller has been an excellent device for providing a flexible platform for the testing of various input and output devices. The Handy Board has also allowed us to test various software configurations. However, there are several serious drawbacks to this device including weight, physical size, memory and computational speed.

Follow-on efforts should also incorporate an expanded electronics capability to include a complete inertial navigation package. This expanded capability will improve ATRP accuracy and responsiveness. At the suggestion of our ONR sponsors, a sensor package that accurately correlates video images with ground coordinates would be developed.

Research and development would also be performed to reduce the size and weight of the ATRP vehicle. The parafoil aerodynamics will be optimized to increase the range and duration of the ATRP flight regime. A study and potentially design and build of a motorized version of the ATRP was also suggested by our ONR visitors.

References

1. GopalaPillai, S., Tian, L., Beal, J., "Detection of Nitrogen Stress in Corn Using Digital Aerial Imaging", 1998.
2. Johannsen, C.J., "Agricultural Applications of Remote Sensing", 2001.
3. Martin, F. G., "The Handy Board Technical Reference", Gleason Research, 2000.
4. Clark, Kevin, "Sampcode.c"
http://www.egr.msu.edu/classes/ece482/Teams/99spr/design5/deliverables/code/main.c
5. Martin, Fred, "Serialio.c" http://216.239.33.100/search?q=cache:Y3ohOvJ5bI4C:www-isl.ece.arizona.edu/~soccer/Fall2002/ai/serialio.pdf+serialio.c+fred+martin&hl=en&ie=UTF-8
6. Mulders, Symen "Comm.c" http://clubs.plattsburgh.edu/csc/handy_car/source/comm.c
7. Drushel, Richard F., "Conio.c" http://handyboard.com/software/contrib/drushel/conio.c
8. Lang, Kam, "RF Serial Communcation for the MIT Handy Board or Motorola 68HC11 MCU"
9. MATLAB, http://www.mathworks.com.
10. DeTurris, D., Ervin J. and Alptekin, S., Development of an Autonomous Tactical Reconnaissance Platform, Proceedings of 17th AIAA Aerodynamic Decelerator Systems Technology Conference, May 20-22, 2003, Monterey, California.
11. Cox, E., The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems. Boston, MA: Academic, 1994.
12. Combs, William E. and James Andrew, Combinatorial Rule Explosion Eliminated by a Fuzzy Rule Configuration, *IEEE Transactions on Fuzzy Systems*, Vol. 6, No. 1, 1998.

13. Dick, S. and Abraham Kandel, A comment on "Combinatorial Rule Explosion Eliminated by a Fuzzy Rule Configuration" by William E. Combs and James Andrews, *IEEE Transactions on Fuzzy Systems*, Vol. 7, No. 4, pp: 475-477, 1999.

14. Mendel, J. M. and Qilian Liang, Comments on Combinatorial Rule Explosion Eliminated by a Fuzzy Rule Configuration, *IEEE Transactions on Fuzzy Systems*, Vol. 7, No. 3, pp: 369-373, 1999.

15. Weinschenk, J. J., W. E. Combs and R. J. Marks II, Avoidance of Rule Explosion by Mapping Fuzzy Systems to a Union Rule Configuration, *FUZZIEEE*, accepted for publication, 2003.

Bibliography

1. Passino, K.M. and Yurkovitch, S., Fuzzy Control, Addison-Wesley, CA, 1998.

2. Sugeno, M., "Development of an Intelligent Unmanned Helicopter", presentation at the World Automation Congress, May 10-14, 1998, Anchorage, Alaska.

3. Swanson, S., Fuzzy Control of the Shuttle Training Aircraft, Applications of Fuzzy Logic Towards High Machine Intelligence Quotient Systems, Prentice Hall, NJ, 1997, pg. 387.

4. Fournell J. and S. E. Alptekin, "Fuzzy Logic Fly-by-Wire System", Proceedings of World Automation Congress, May 10-14, 1998, Anchorage, Alaska.

5. Ervin J. and S. E. Alptekin, "Fuzzy Logic Control of a Model Airplane", Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics, October 11-14, 1998, San Diego, California.

6. Alptekin, S.E., Fournell, J. and Ervin, J. "An Investigation of Fuzzy Logic Control of a Complex Mechatronic Device", Proceedings of the 2nd International Conference on Recent Advances in Mechatronics, ICRAM'99, May 24-26, 1999, Istanbul, Turkey.

7. Wind Dance Parafoil Kite http://www.seattleairgear.com/

8. Guided Parafoil Air Delivery System–Light, http://www.fas.org/man/dod-01/sys/ac/equip/gpads-l.htm

9. Learning and soft computing : support vector machines, neural networks, and fuzzy logic models , Kecman, V. (Vojislav), Cambridge, Mass.: MIT Press, c2001.

10. Advanced fuzzy systems design and applications, Yaochu Jin, Heidelberg ; New York : Physica-Verlag, 2003.

11. Soft computing: integrating evolutionary, neural, and fuzzy systems, Andrea Tettamanzi, Marco Tomassini. Berlin; New York : Springer, c2001.

12. Fuzzy logic techniques for autonomous vehicle navigation, Dimiter Driankov, Alessandro Saffiotti, editors.Heidelberg ; New York : Physica-Verlag, c2001.

13. Intelligent control systems using soft computing methodologies, edited by Ali Zilouchian, Mo Jamshidi. Boca Raton, Fla.: CRC Press, c2001.

14. Soft computing agents: new trends for designing autonomous systems, Vincenzo Loia, Salvatore Sessa, editors.Heidelberg ; New York: Physica-Verlag, c2001.

15. Introduction to neuro-fuzzy systems / Robert Fuller, Heidelberg ; New York: Physica-Verlag, c2000.

16. Soft computing in mechatronics, Kaoru Hirota, Toshio Fukuda (eds.), Published: New York : Physica-Verlag, c1999.

17. Soft computing for intelligent robotic systems, Lakhmi C. Jain, Toshio Fukuda, (eds.), Heidelberg; New York : Physica-Verlag, c1998.

22

Fast, Remote, Infrared Spectroscopy of a Rocket Plume

Brian Higgins

Mechanical Engineering Department
Cal Poly State University

Fast, Remote, Infrared Spectroscopy of a Rocket Plume

C3RP Final Report

December 2003

Brian Higgins

Table of Contents

List of Figures

List of Tables

Glossary

COTS   commercial off-the-shelf
CT   computer tomography
HeNe   helium-neon laser
IR   infrared
LOS   line-of-sight
MR   Oxidizer-to-fuel mass mixture ratio
NPS   Naval Postgraduate School, Monterey, CA
P_o rated laser power in Watts
sr   steradian
SSI   Spectral Sciences Inc., Burlington, MA
STIRS   Sierra tomographic infrared spectroscopy
TDL   tunable diode laser
TDLAS   tunable diode laser absorption spectroscopy
λ wavelength of light, typically in μm
Φ fuel-air equivalence ratio (relative to stoichiometric)

v

1.0 INTRODUCTION

The goal of this proposal was to integrate existing optical techniques to measure temperature and species concentrations in a rocket plume. The experimental work load was done at Cal Poly and Sierra Engineering Inc provided technical assistance for the analysis.

In order to simultaneously measure temperature and species concentrations, infrared absorption spectroscopy is combined with infrared emission spectroscopy. Each of these techniques is line-of-sight measurement. By performing line-of-sight measurements at many cross sections of the rocket plume, tomographic inversion can be used to acquire spatially resolved data.

We choose to use infrared fiber-optics cables and transfer optics. While fiber optics add a certain level of complexity, they also enable remote location of the spectrometer and light source, allowing for quicker rotation and translation of the collection optics necessary to capture sufficient cross sectional data in the short time allotted for a rocket test (typically under ten seconds). Further, fiber optics allow increased portability and adaptability to larger and different combustion flow streams (e.g., power plant exhaust stacks or automotive exhaust pipes).

Dr. Brian Higgins was the principal investigator (PI) for the work performed at Cal Poly and Dr. Jeffrey Muss was the PI for the work performed at Sierra Engineering Inc. Mr. Dowdy did much of the optical design and experimental work as part of his Masters Thesis research. Mr. Dowdy's thesis is scheduled to be defended in the coming weeks. Dr. Muss, Mr. Dowdy, and I are indebted to ONR and C3RP for the opportunity to extend our work in this field and we wish to express our mutual appreciation for the C3RP program.

2.0 SUMMARY OF PROGRAM FINDINGS

The following is a summary of program's findings:

1. Infrared fiber optic collection can be effectively integrated with the Spectraline ES-100 spectrometer. Inconsistencies in the fiber cleaving process and the critical nature of alignment of external components, e.g. coupling lenses, makes incorporation of a back-illumination source within the ES-100 optical path highly desirable.

2. The selected 800 μm diameter multi-mode C3 optical fiber possesses several undesirable characteristics:

   - the relatively large diameter results in poor focusing characteristics;

   - the large diameter and multi-mode nature results in smearing of spectral energy across several pixels centered around the incident wavelength;

   - the relatively large diameter and multi-mode nature of the fiber make the fiber transmissivity sensitive to changes in the fiber bend radius;

   - the strong fiber attenuation between 3.95 and 4.35 μm necessitates significant correction to CO_2 emission measurements.

3. It is extremely difficult, if not impossible, to efficiently couple broadband IR radiation into a fiber and re-transmit it as a collimated beam. This is a fatal flaw for a broadband absorption diagnostic.

4. In spite of the problems with absorption measurements and the fiber, a <u>stationary</u> broadband fiber-coupled emission diagnostic appears to be practical using the subject fiber.

5. Low cost fiber-coupled tunable diode laser (TDL) absorption spectroscopy (TDLAS) diagnostics exist that could be used to simultaneously determine water concentration and temperature in rocket plumes. The system is based on commercial off-the-shelf (COTS) telecommunication equipment and does not require active cooling. It appears that these systems can be multiplexed to provide measurements from multiple viewing locations at a cost of approximately $4500 per channel. Effort is still needed to deconvolve the measurements to create spatial distributions of temperature and concentration.

6. TDLAS technology may be extendable to measurement of CO_2 and CO using similar COTS telecommunication equipment.

8

3.0 REVIEW OF TECHNICAL RESULTS

3.1   Development of a Fiber-Coupled Diagnostic

Cal Poly led the integration of the fiber-optic transmission and collection components. The basis point was the Sierra Tomographic Infrared Spectroscopy (STIRS) system. The STIRS system consists of a cavity blackbody, a Spectraline[1] ES-100 mid-IR spectrometer and traversing optics (Figure 1).

The initial goal of this project was to develop the flexibility necessary to take infrared emission-absorption measurements from several viewing angles simultaneously. The hope was that fiber-optic coupling would afford this flexibility, permitting several of the new STIRS setups to be located <u>near</u> a rocket test stand.

Requirements for fiber optic integration are:

1. The fiber optic collection optics must be compatible with the Spectraline ES-100 mid-IR spectrometer.

2. The system have good spectral transmission characteristics between 2.0 and 5.0 μm, with particular interest in the CO_2/H_2O band located between 2.3 and 2.7 μm, the C-H stretch located around 3.4 μm and the CO_2 band between 4.0 and 4.6 μm.

3. The resultant system should permit the ES-100 and the IR source, used for transmission measurements, to be located remotely, with the fiber optic ends traversed.

4. The free path between fiber ends, e.g. between the output end of the blackbody transfer optics and the input end of the spectrometer optics, be between 3 and 30 inches (the GAP in Figure 2).
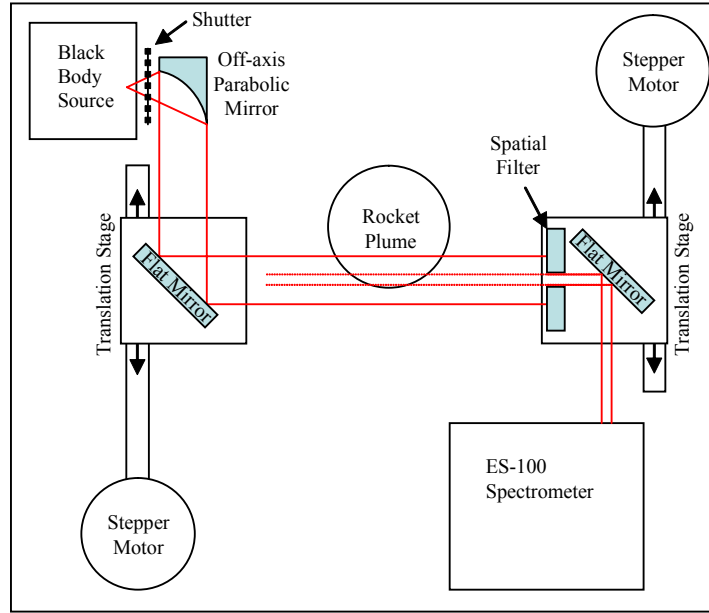
Figure 1: Schematic of STIRS System

An 800 µm diameter Chalcogenide C3 fiber produced by REFLEX Analytical Corp.[2] was selected based on transmission characteristics, core diameter, price and availability. These and other design selections were reviewed early in the project, with Drs. Chris Brophy and Jose Sinibaldi of the Naval Postgraduate School (NPS) assisting Dr. Muss in the critical review process.

3.1.1   Fiber System Characterization

The path the light takes from the blackbody source to the detector follows four connections (see Figure 2). The connections are, in order:

The path the light takes from the blackbody source to the detector follows four connections (see Figure 2, Figure 5 and Figure 6). These connections have been further refined and optimized as the described in the subsequent subsections. The connections are, in order:

- Point 1: the optical connection from the blackbody source to the first fiber.

- Point 2: the termination of the first fiber, directing the light across the plume.

- Point 3: the collection optics on the other side of the plume into the second fiber.

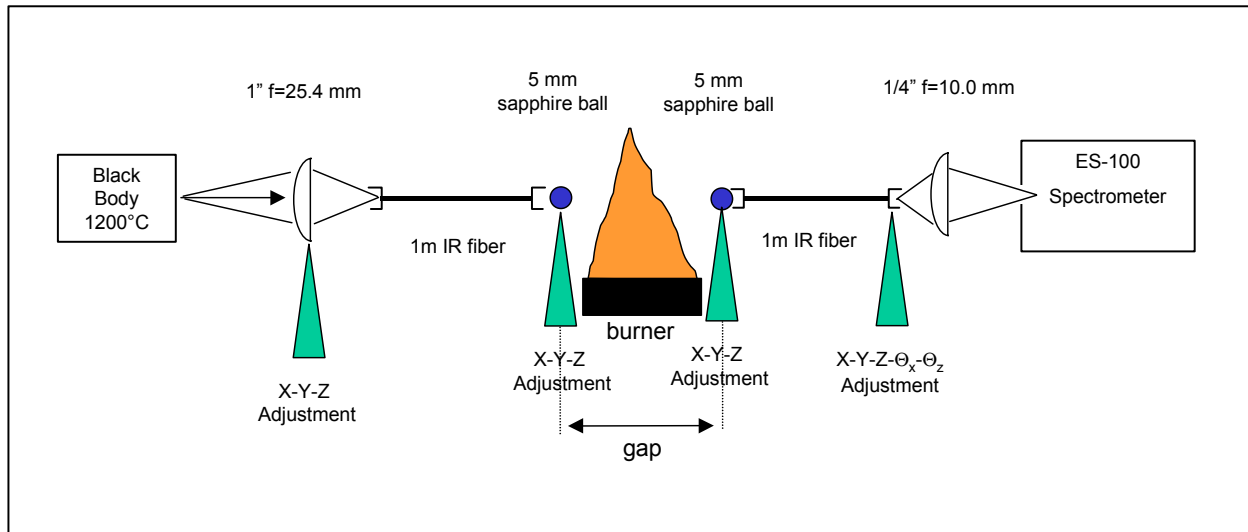- Point 4: the optical connection from the second fiber into the detector.

Figure 2: Schematic of Fiber Optic Test Configuration

Point 1: Blackbody Source Connection

The final configuration evaluated used a 25.4-mm plano convex sapphire lens with a 25.4-mm focal length to focus light from the cavity blackbody source into an IR fiber (Point 1). The lens-fiber assembly has two degrees of freedom relative to the blackbody (course adjustment) and there are three linear degrees of freedom between the lens and fiber. Due to the one-inch cavity of the blackbody, this is ample. A photograph of the assembly is included as Figure 3.

Initial efforts used an available smaller 0.25-inch diameter cavity blackbody with a maximum temperature of 1050 ºC. Preliminary testing using a fiber coupled directly to the spectrometer indicated that the available energy was marginal at best. The larger 1-inch diameter blackbody source with a maximum temperature of 1200 ºC was borrowed from the Naval Postgraduate School (NPS). This more than doubled the spectral power throughput of the system.

The initial system design also used sapphire ball lenses to couple between the blackbody source and the fiber. Early tests showed that they did not adequately couple the light with the fiber optic. The small plano-convex sapphire lens produced good coupling (20x improvement) of the IR light into the fiber and out into the spectrometer. The poor coupling efficiency of the ball lens was attributed to numerical aperture (NA) mismatch.

Point 2 and Point 3: Ball Lens Connection to Fiber on Either Side of the Gap

Two 5-mm sapphire spherical ball lenses are used to collect and collimate light from the fiber, transmit the collimated beam across the free-space gap (rocket plume), recollect the light and couple it back into the fiber. The fibers are mounted at the back focal length of each ball providing the proper gap between the edge of the ball and the end of the fiber. Each ball has three degrees of adjustment (x, y, and z) with regards to the fiber; thus, assuring maximum coupling efficiencies.

11

Figure 3: Side View of Blackbody Connection Assembly (Point 1)

The collection and collimation of the light exiting the fiber, prior to "jumping the gap", is the major shortcoming of this fiber optic integration. Poor collimation of the light results in a significant drop in the radiation collected on the opposite side of the plume, and a precipitous drop in signal strength for the absorption measurements. A significant effort was spent trying to improve the collimation of the radiation exiting the fiber prior to jumping the gap. Plano-convex lenses were evaluated alone and in conjunction with the sapphire ball lenses. None of these variations yielded a significant improvement in signal strength. Several sources for this problem have been hypothesized:

- The fiber diameter (0.8 mm) is relatively large. Therefore exiting light does not appear to the ball lens (5 mm diameter) to be emanating from a point source;

- Since index of refraction changes with wavelength, the focal length of the lens is variable, complicating collimation of broadband blackbody radiation (aberrations);

- Uncertain and uncontrolled cleaving conditions at the end of the fiber further exacerbate difficulties in collecting the radiation exiting from the fiber with a simple lens.

Use of an achromatic doublet may have improved collimation some, but no sapphire lenses were readily available.

Point 2 to Point 3: "Jumping the Gap"

In order to work as a plume diagnostic there must be a physical gap between points 2 and 3. We have varied this distance from zero to three inches. Doing this we have found a $1/(4X^2)$ relationship (Figure 7). That is, for each doubled of distance across the gap, the net throughput into the spectrometer goes down by a factor of four.

Point 4: Spectrometer connection

The coupling from the second fiber to the spectrometer includes adjustments with five degrees of-freedom (3 linear and 2 rotational). The fiber coupling system mounts to the outer attachment ring of the ES-100 that houses the entrance slit. A solid model and picture of this assembly are included as Figure 4.
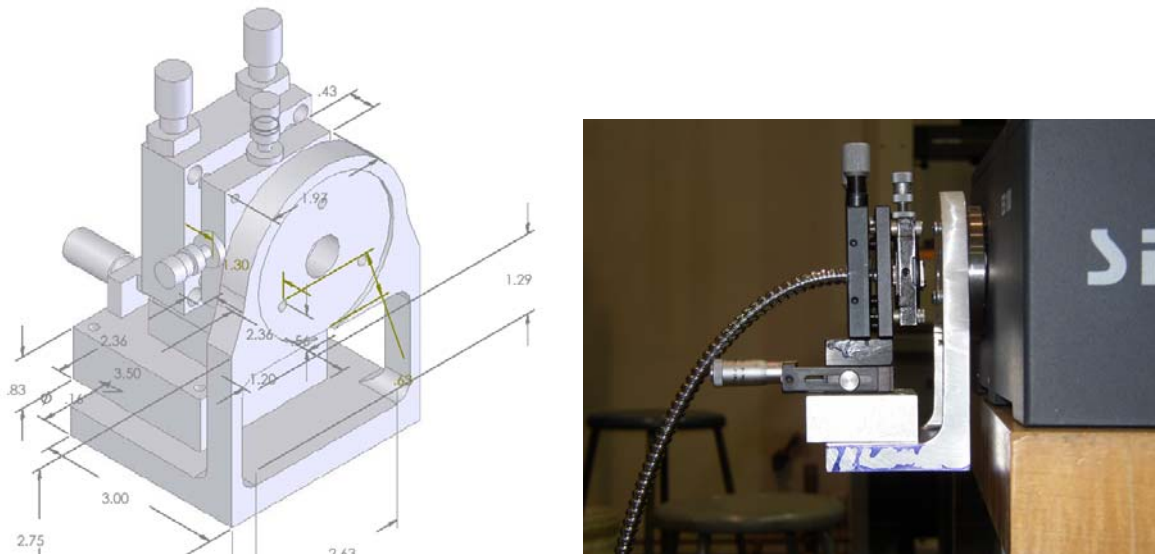

Figure 4: Spectrometer Fiber Connection Assembly, Model (L) and View Showing Attachment to ES-100 (R)

The purpose of these adjustments was to provide the maximum intensity to the ES-100 spectrometer. The goal was to fill the first internal mirror in the ES-100 (¾-inch, 180mm away) with radiation transmitted by the fiber and focused on the 0.6 mm wide slit. A 10-mm F 1.0 lens (10mm dia, 10mm focal length) was fixed 1.779 inches in front of the entrance slit. The distance of the slit to the mirror is approximately 6 inches. This equates to minimum lens diameter of 5.65 mm, less than the 10 mm of the selected lens. The fiber end must be located at the middle or centerline of the z-axis adjustment. Using lens the focal length diameter and axial off-set between the lens and slit, the distance from the fiber tip to the lens needs to be 0.506±0.002 inches off the centerline. The axial (z-axis) translation stage was set to locate the fiber tip 0.506 inches away from the lens, and allow an adjustment of ± 0.25 inches.

13

The fiber mounting unit also includes an X-Y adjustment stage mounted to a two axis tilt, which in turn is mounted to the axial translation stage. The whole lens assembly is mounted to an L-shaped bracket that connects the fiber adjustment system to the spectrometer-mounting flange. This assembly mounts in-line with the centerline of the optical axis of the spectrometer.

3.1.2 Scaling of Available Power

The system throughput and loss characteristics were evaluated using data for the individual segments. The blackbody emissive power was recalculated to be 89.3 kW/m^2-um at 1.97 μm from an ideal 1200 °C source. Dividing by π steradians (sr) results in 28.42 kW/m^2-sr-um radiated power, a number that can be directly compared to the radiation detected by the ES-100 spectrometer.
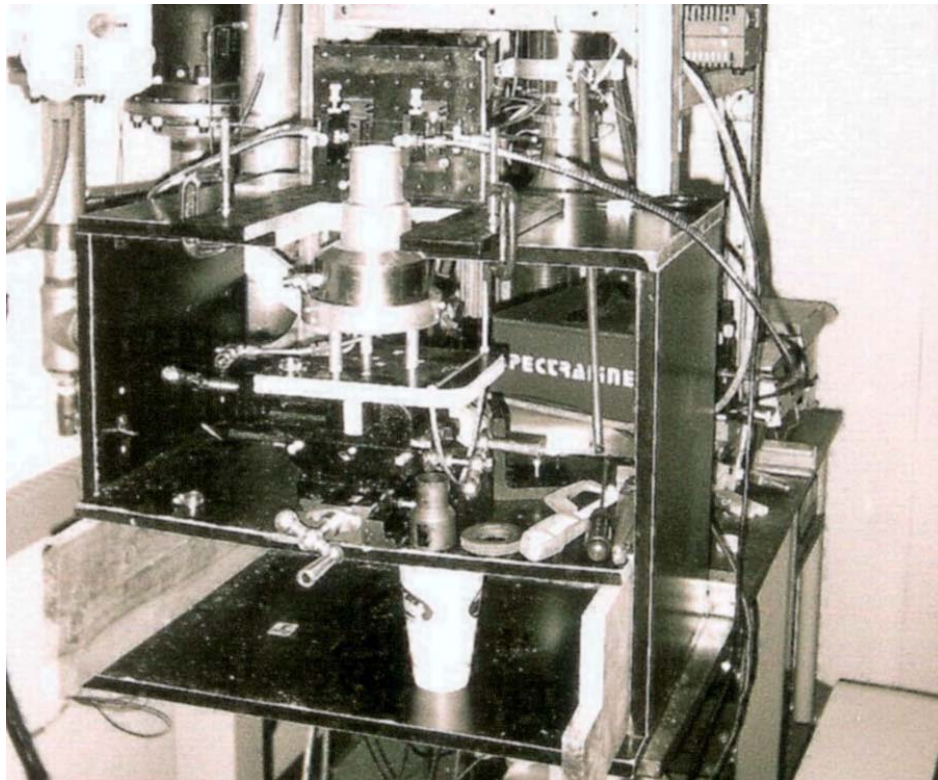

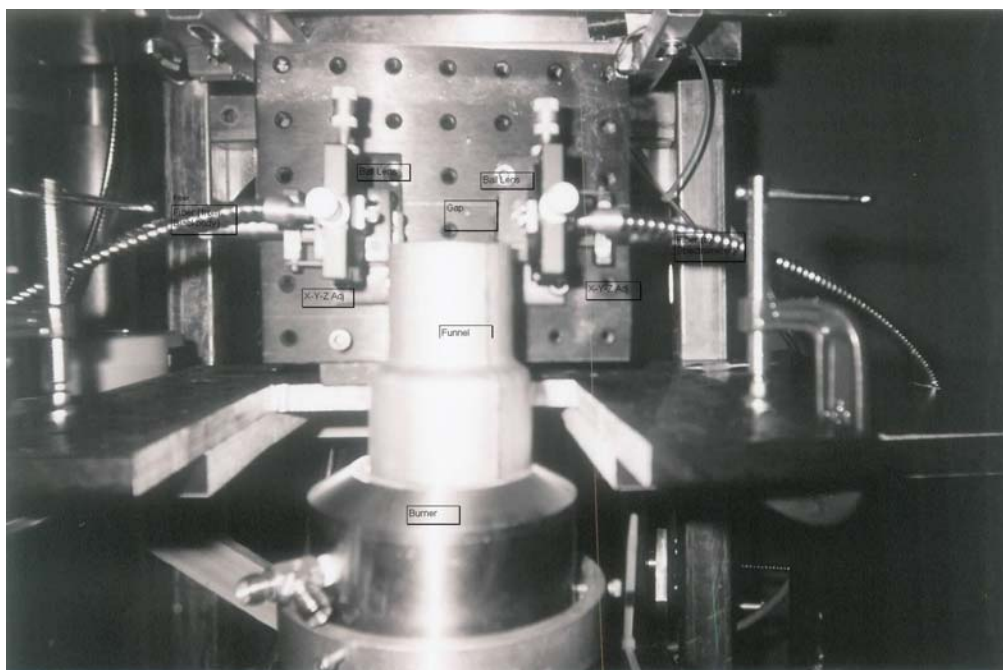**Figure 5: Photo of Fiber Optic Test Setup**

14

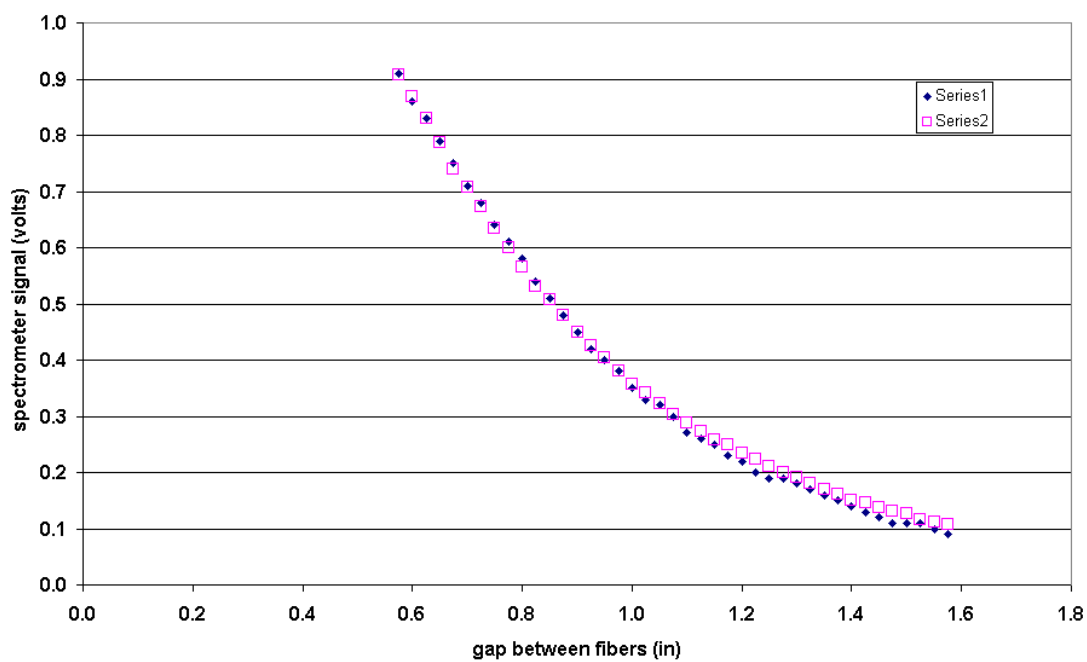**Figure 6: Close-up of Free-Space Fiber Optic Test Setup**

Figure 7: Effect of Free-Space Gap on Transmitted Spectral Signal

15

The initial configuration tested used a fiber to directly connect the black body to the spectrometer. The test configuration used the two sapphire plano-convex lenses depicted in Figure 2, to couple the radiation from the black body into the fiber and out of the fiber onto the slit of the spectrometer. The detected radiation was 12.3 kW/m^2-sr-μm, corresponding to a 43.5% coupling efficiency or 3.62 dB. This efficiency agrees well with typical component data –

- transmission through each Sapphire lens $\cong$ 80% (0.969 dB)

- attenuation through fiber cable at 0.2 dB/m (Figure 8) $\cong$ 95.5%

- fiber coupling efficiency, including Fresnel losses, at 0.5 dB/connection $\cong$ 89.1%

resulting in a 50.6% theoretical coupling efficiency. The coupling efficiency from the black body into the fiber by the sapphire lens should be similar to the coupling efficiency from the fiber onto the spectrometer slit by the sapphire lens. Using the fiber attenuation as an "absolute", the efficiency at each end of the optical train (points 1 and 4) is 67.5% or 1.71 dB.

When the optics train is fully optimized with no gap (first ball lens in contact with the other ball lens), the measured radiation was 3.98 kW/m^2-sr um, 14% of the available power. The coupling efficiency of the fiber, through the two ball lenses in contact with each other, and back into the fiber was calculated to be 33.8%, or 58.1% for each ball lens (2.36 dB), including the fiber-ball lens connection. This efficiency is consistent with published values of 55-60%.
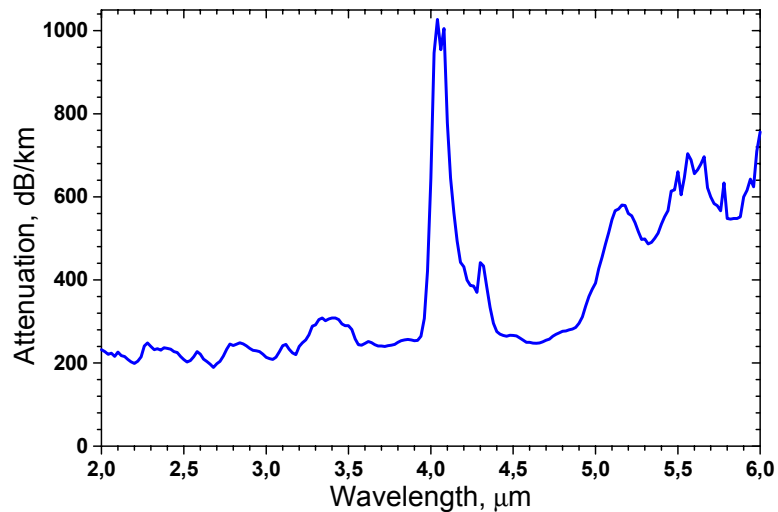

Figure 8: Manufacture's Attenuation Curve for C3 Fiber[2]

In order to maximize signal-to-noise characteristics and yet have reasonable headroom for the emission signal above the pure blackbody signal, the maximum blackbody signal detected by the ES-100 needs to be about 4 volts or 8 kW/m^2-sr-μm. Applying the total system efficiency of 14% (neglecting the effect of the gap for now, assuming ball lens touching ball lens), 57.2 kW/m^2-sr-um or 180 kW/m^2-um is required from the black body. This corresponds to a blackbody temperature of approximately 1500 °C.

Now look at a fiber optic system that includes a 5-3/4 inch gap. In order to achieve the present 3.98 kW/m^2-sr-μm at the spectrometer, the input to the second ball lens (point 3) must be 10.64 kW/m^2-sr-μm. The attenuation across the a 5-3/4 inch gap calculates as 1/(4 X^2), corresponding to 1404 kW/m^2-sr-μm exiting the first ball lens. The attenuation through the upstream portion of the fiber optic system is also 4.27 dB for a 1 meter fiber, requiring a black body output of 3753 kW/m^2-sr-μm. This equates to a black body temperature of 3028 °C. well above typical black body temperatures.

Therefore, another energy source is required.

3.1.3   Evaluation of a Laser Source

The blackbody source has now been shown experimentally and analytically to have insufficient optical throughput for a practical fiber-based transmission diagnostic. A variety of "gap jumping" optical component configurations were investigated in an effort to enhance collimation and throughput, but without major improvements.

An IR laser (3.79 μm) was used to assess potential improvements offered by coherent, narrow wavelength sources. The laser source was positioned at point 1-2. It illuminated the existing blackbody 1-inch plano-convex lens. The fiber is connected to the focus point of the lens. The light is transmitted through the fiber, then to the existing lens assembly at the spectrometer slit (Figure 9). It was difficult to align the laser with the ball-fiber coupling. Alignment was easier with the 1-inch lens assembly, as opposed to the existing 5mm ball lens assemblies. The laser requires more precise x-y-z alignment and at least 2 axis tilt adjustment. Misalignment was intentionally used during these experiments to attenuate the power ultimately detected by the spectrometer. The distance from the laser exit to the lens was varied from 6 to 30 inches, with the alignment adjusted to keep the detected power below the spectrometer's maximum (about 20 kW/m^2-sr-μm). Even at a separation distance of 30 inches an intensity of 10 kW/m^2-sr-μm was achievable. The acceptance characteristics of the fiber make the detected intensity very sensitive to alignment, dropping to near zero with small adjustments.

A 5 mW continuous output helium-neon (HeNe) laser operating at 3.79 μm was borrowed from UC Berkeley in order to evaluate the effectiveness of coherent light to improving optical throughput. The power density of the laser operating at 3.79μm was calculated to be approximately 277,500 kW/m^2-sr using the following equation[3]:

$$I_o = 2 \times 10^{12} \, P_o/(\pi\lambda^2) \qquad\qquad (1)$$

where the rated laser power P_o is in Watts, and the wavelength λ is in μm. This is about 75 times more power than the previous calculations showed were needed to produce a feasible system with a 6-inch gap.

The laser was tested through a flame with the laser positioned 6 inches from the focusing lens, as depicted in Figure 9. The detected peak intensity at the spectrometer ranged between 8-10 kW/m^2-sr-um. These tests showed that significant power could be coupled into a fiber-coupled spectrometer. However, there appear to be much better technical solutions if discrete wavelength radiation is to be used for emission or absorption diagnostics.
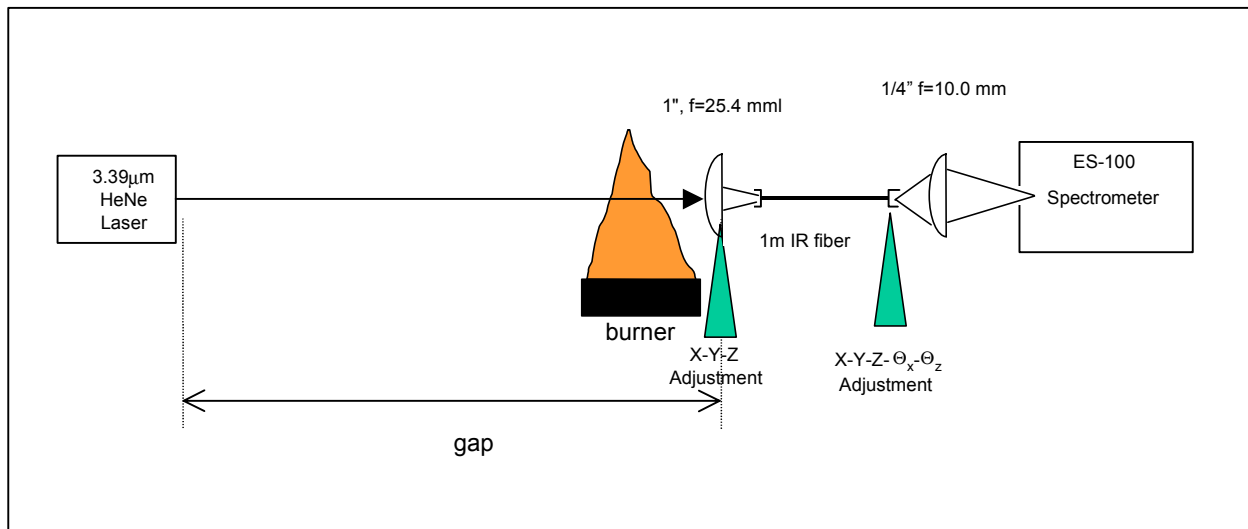
Figure 9: Schematic of Test Configuration using IR HeNe Laser

3.2   Experimental Evaluation of the Diagnostic

A series of evaluation tests were performed over a uniform flame. This section describes the setup and the experimental results.

3.2.1   Burner Design

A key to validation of any emission-absorption diagnostic is comparison of measurements to theoretical results. The first-level validation requires measurements in a well-controlled, uniform flame. Dr. Robert Cattolica, formerly of Sandia National Labs and currently a professor at UC San Diego was contracted to design a calibration burner for this purpose. The 50-mm. diameter premixed burner design is based on an earlier 1-inch diameter version previously used at Sandia National Labs. The burner produces for lifted flat-flame operation with premixed methane and air. The burner is constructed from micro-capillary tubes (about 1200 - 0.058" OD/0.048" ID) and includes a nitrogen shroud (Figure 10). Fabrication costs were estimated at $3000.

Figure 10: Assembly Drawing of 50-mm Premixed Burner (p/n SEI00331)

19
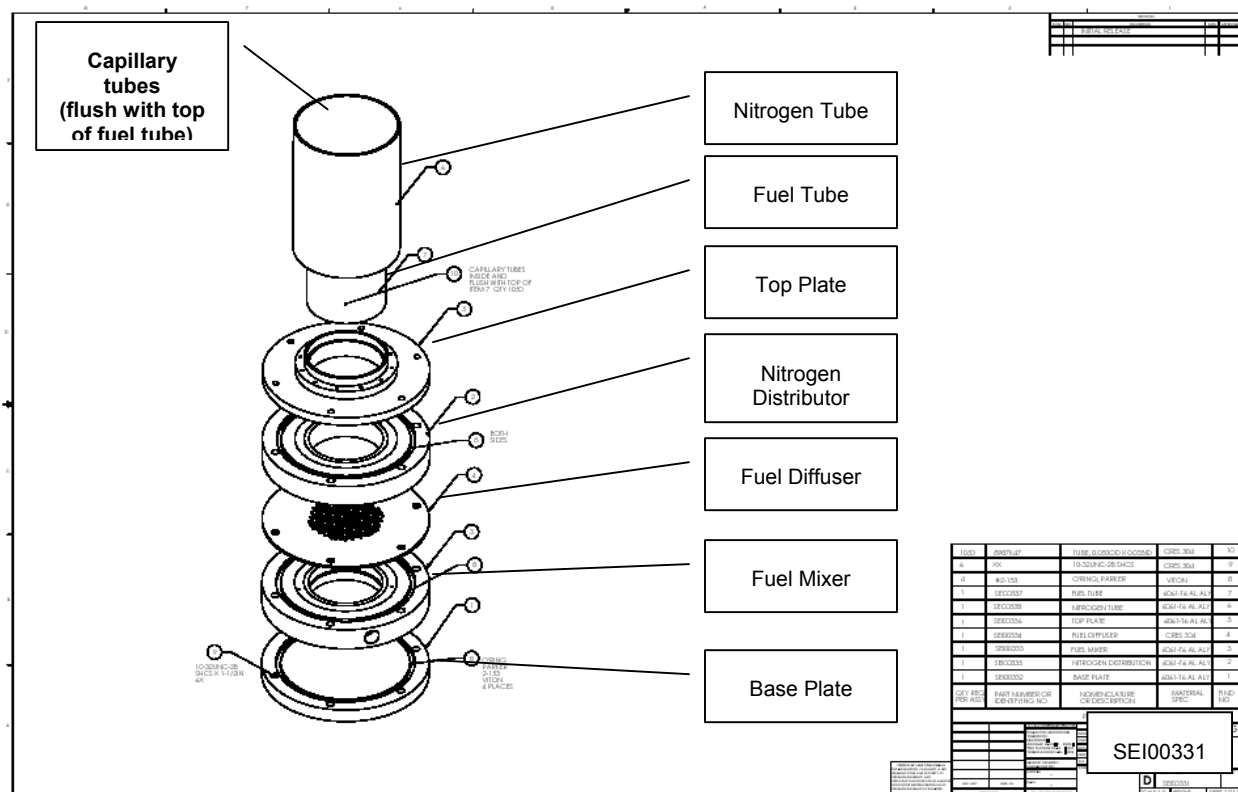
A decision was made to conserve project funds and we borrowed a McKenna burner from Lawrence Berkeley National Laboratories for the hardware evaluation phase of the project. A McKenna burner is a water-cooled sintered-disk burner that produces a uniform premixed flame anchored to the burner surface. Two neck-down segments were constructed to reduce the burner exhaust diameter from 3.0 to 1.125 and 1.90 inches (see Figure 11).

3.2.2   Experimental Setup

The optical train described in Section 3.1.1 was assembled in the combustion laboratory at Cal Poly. The optical stages that secure the "free" ends of the fiber optics were mounted to a fixed optical base (translation stages were retained to move the fiber ends up/down and adjust the separation gap). The burner was mounted to a 2-axis translation system providing x and y-axis adjustment for the burner. Each turn of the handle (360°) moves the burner a linear distance of 1/8–inch. In this way, the optics remained stationary and the burner was traversed through the optical path (Figure 11).
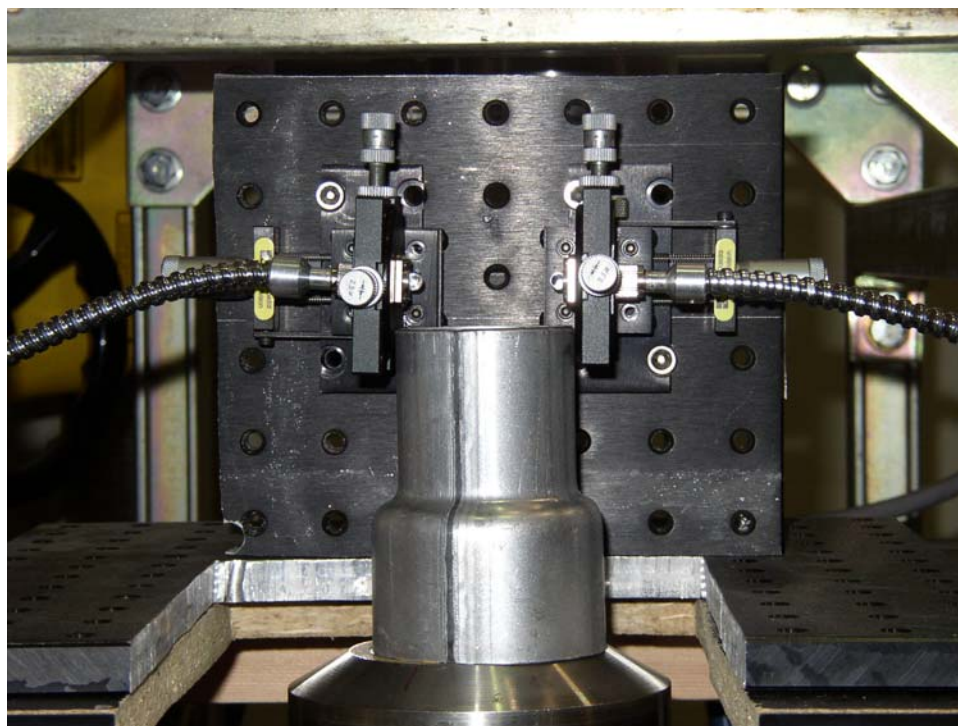

Figure 11: "Free Ends" of Fiber System (points 2 and 3) Mounted over Traversing Burner, with 1.90-inch Neck-down Adapter Mounted on Burner.

A gas panel was constructed to monitor fuel and air flows into the burner. Due to a lack of resolution in the flow meters, the stoichiometry reported below was estimated based on chemiluminescence observations of the flame.

20

3.2.3   Test Descriptions

A total of six (6) test series were conducted. Each series consists of a number of parallel projections across the flame, with the optics stationary and the burner traversed. The data for each parallel projection included 6000 spectral scans of purely radiative emission and an additional 6000 scans where the signal was the radiative emission convolved with the transmission characteristics of an infrared (IR) source on the opposite side of the flame. Two IR sources were tested, a cavity black body operating at 1200 ºC (series AA through AD) and a helium-neon (HeNe) laser operating at 3.79 μm (series AF and AG).

Test series AA through AD were performed using methane-air while series AF and AG used propane and air as the reactants. All flames were premixed, generated using a 3.0 inch diameter sintered disk (McKenna) burner. Table 1 lists the infrared source, flame diameter and approximate fuel-air equivalence ratio Φ for the six test cases. Series AA through AD used flow area reducers to converge the flow and produce smaller burner diameters. Series AF and AG used the full burner face diameter.

Temperatures were measured at several points along the radius as shown in Figure 12. These data have not been corrected for radiation or conduction, so the true temperature could easily be several hundred degrees higher than reported.

There are marked differences in data amplitude for adjacent pixels as a result of the odd/even pixel arrangement in the ES-100 spectrometer (see Figure 14). This study continued the practice of only using test data for the even pixels for tomographic data reduction[4].

Table 1: Test Series Descriptions

SERIES	IR SOURCE	EXIT DIAMETER (IN)	Φ
AA	Black Body, 1200 ºC	1.90	1.0
AB	Black Body, 1200 ºC	1.90	1.2
AC	Black Body, 1200 ºC	1.125	0.8
AD	Black Body, 1200 ºC	1.125	1.2
AF	HeNe Laser, 3.79um	3.0	1.0
AG	HeNe Laser, 3.79um	3.0	1.0

21
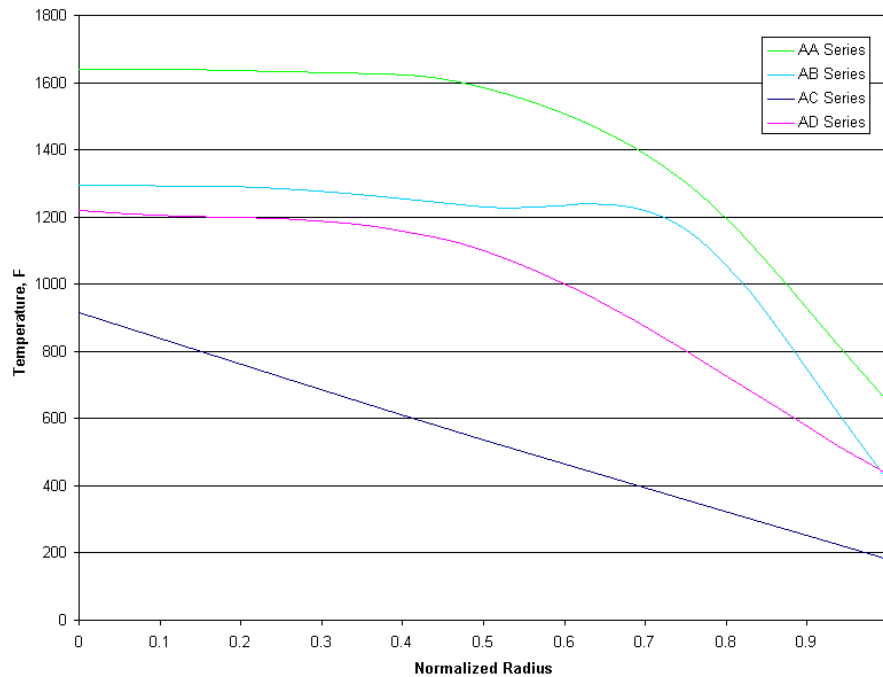
Figure 12: **Temperature Profile Across Burner Centerline for Series AA through AD.**

3.2.4   Data Reduction Procedure

Since the test data were taken across an axisymmetric burner, data was post-processed for deconvolution with Sierra's FAST onion-peeling tomographic deconvolution software. The FAST deconvolution code uses a series of parallel line-of-sight (LOS) transmission (τ) and emission (N) measurements to determine temperature and concentration for the major combustion species. This section documents how these LOS values are calculated from raw spectroscopic data.

Raw Data Manipulation

The first step required is to isolate the individual scans for each station, then average these data together for each pixel in the detector. When finished, the data can have any combination of the three following attributes:

- Emission-Only <u>or</u> Emission-and-Transmission

- Combustion Occurring <u>or</u> No Combustion Occurring

- Odd or Even pixels

Since odd-pixel data are not used during data reduction, four outcomes are possible:

22

- N_o = Emission-Only with No Combustion (background emission radiation)

- $N_o + I_o$ = Emission-and-Transmission with No Combustion (background blackbody and emission radiation)

- N_i = Emission-Only with Combustion (emission radiation)

- $N_i + I_i$ = Emission-and-Transmission with Combustion (blackbody and emission radiation)

Using these outcomes, transmission (τ) and emission (N) are defined as:

$$N = N_i - N_o \tag{2}$$

and

$$\tau = I_i / I_o \tag{3}$$

which gives

$$\tau = [(I_i + N_i) - N_i] / [(I_o + N_o) - N_o] \tag{4}$$

These calculations are demonstrated in the following figures for Case AD, the smallest flame operating fuel-rich. Note that the above calculations are completed for each pixel and at each station. All of these data are simultaneously plotted on the following graphs. The legends represent the location (0 to 1.27) in centimeters from the edge of the burner and 1.27 cm at the burner centerline.The numbers in the plot legend indicates the scan number. Scan 1 is located approximately 0.125 inches outside the burner, with subsequent scans occurring at 0.125 inch intervals. Scan 6 is approximately the burner centerline and Scan 11 is approximately 0.125 inches beyond the far edge of the burner. Unfortunately the data for Scan 6 was lost during Series AD, so Series 7 is used to approximate the burner centerline.

Figure 13 plots the net transmitted radiation as a function of scan location. This plot results from subtracting the mean emission spectra from the mean emission plus transmission spectra at each scan location. The transmission (I/Io) is calculated by pixel-wise dividing this data by the mean black body radiation curve without the flame present (Io). Only data for the even pixels is presented. The plot also includes a scaled intensity curve for a 1200 °C black body, for reference. It should be noted that the theoretical black body curve peaks at 1.97 μm, just below the lowest wavelength detected by the ES-100.

23

Figure 13 illustrates several interesting and potentially problematic features. First, the data shown has not been corrected for the transmissive characteristics of the optical fiber. The strong drops in signal at 3.2, 3.7 and 4.0 μm correspond with typical attenuation curves for the C3 fibers (Figure 8). The omission of this correction does not effect subsequent transmission calculations, since all final data is ratios. Correction for the absolute emission data includes correction for the spectral transmission characteristics of both the fiber and the sapphire optical components. The second, troubling feature is the large intensity peaks near 4.4 μm for the two near-edge scans (Scans 9 and 10). The emission only data (Figure 15) shows a reasonable progression of intensity across the flame in that wavelength region, although some asymmetry is indicated by the fact that Scans 9 and 10 are lower than their counterpart Scans 3 and 2, respectively. Review of the raw combined emission-transmitted data shows better agreement between Scans 2 and 10 (at the opposite edges of the burner) and good agreement between Scans 3 and 9 (about 0.125 inches inboard of the burner edge). These facts imply that the high transmitted radiation data for Scans 9 and 10 result from problems with the mean emission data for those scans rather than the combined emission-transmission scans; however, this also identifies strong sensitivities that must be accommodated when data subtractions are required.

Emission only data is presented for Series AD in Figure 14 and Figure 15. The effects of the alternating odd and even pixels are obvious in Figure 14. The centerline emission spectra is compared to a RADCAL prediction. The predicted spectra assumes equilibrium species concentrations for a Φ=1.2 methane-air flame, while the temperature has been set to 1200 °F based on the thermocouple measurements. As noted above, the thermocouple measurements have not been corrected for radiation or conduction, nor has the emission data been corrected to account for the fiber's spectral transmission characteristics. The intensity of the plotted RADCAL spectra has been decremented 25% to better match the intensity of the scan data. However, based on the optical throughput tests discussed above, it was expected that the RADCAL data would need to be decremented 72% to account for the optical efficiencies between points 3 and the detector. Emission data for the other three test series also show reasonable agreement with the predicted RADCAL emission intensity (see Figure 19 through Figure 21), indicating that a fiber-coupled emission-only diagnostic may be practical.

24

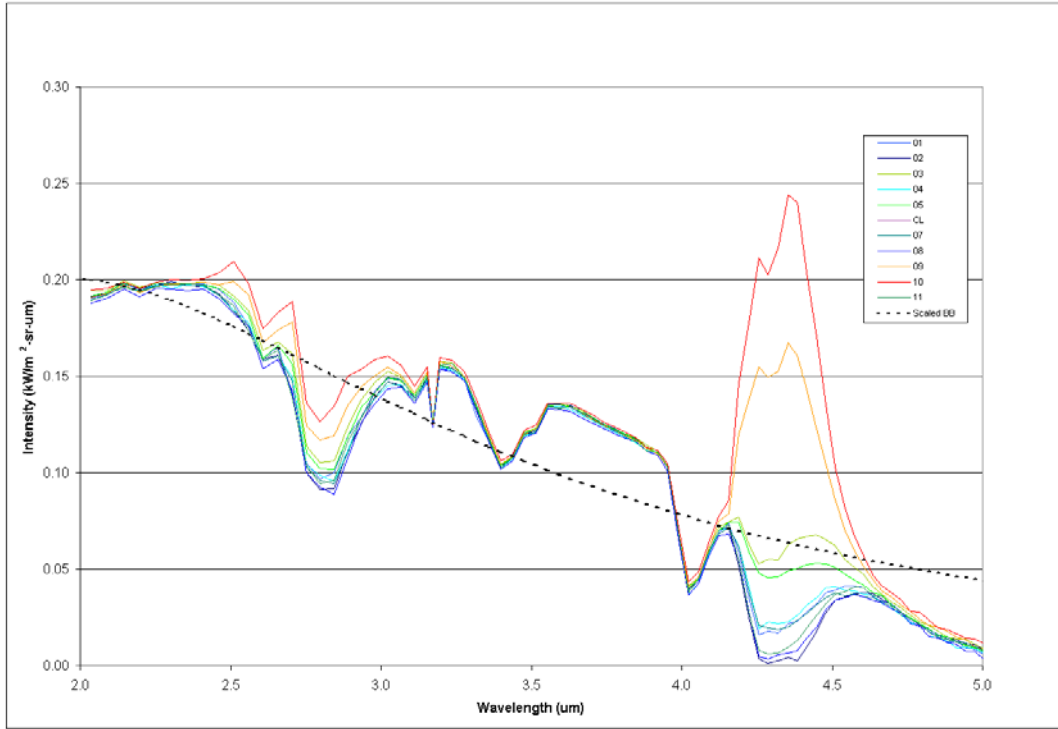**Figure 13: Net Transmitted Radiation for Series AD for Even Pixels at All Measurement Stations.**
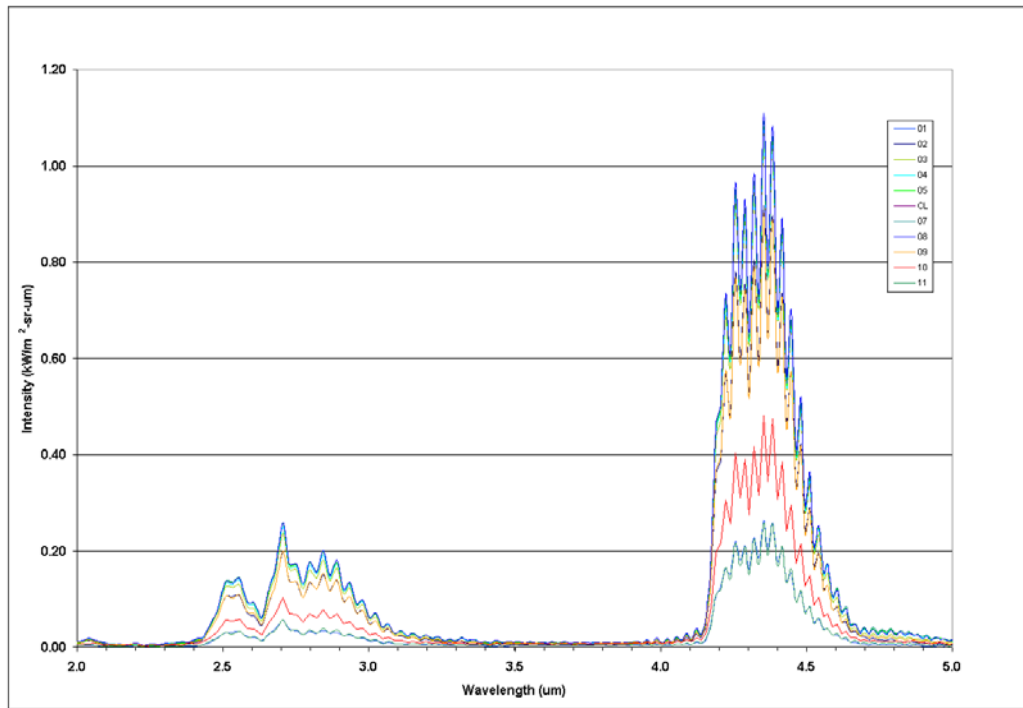

Figure 14: Emission-Only Data for All Pixels as a Function of Scan Position for Series AD.

25
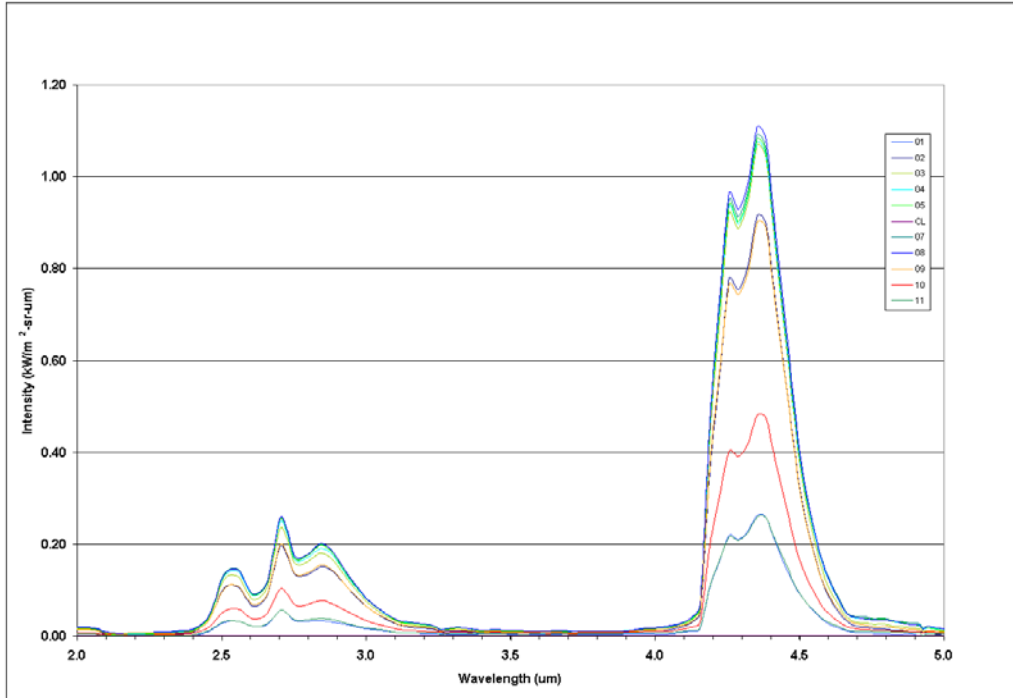
Figure 15: Emission-Only Data for Even Pixels as a Function of Scan Position for Series AD.
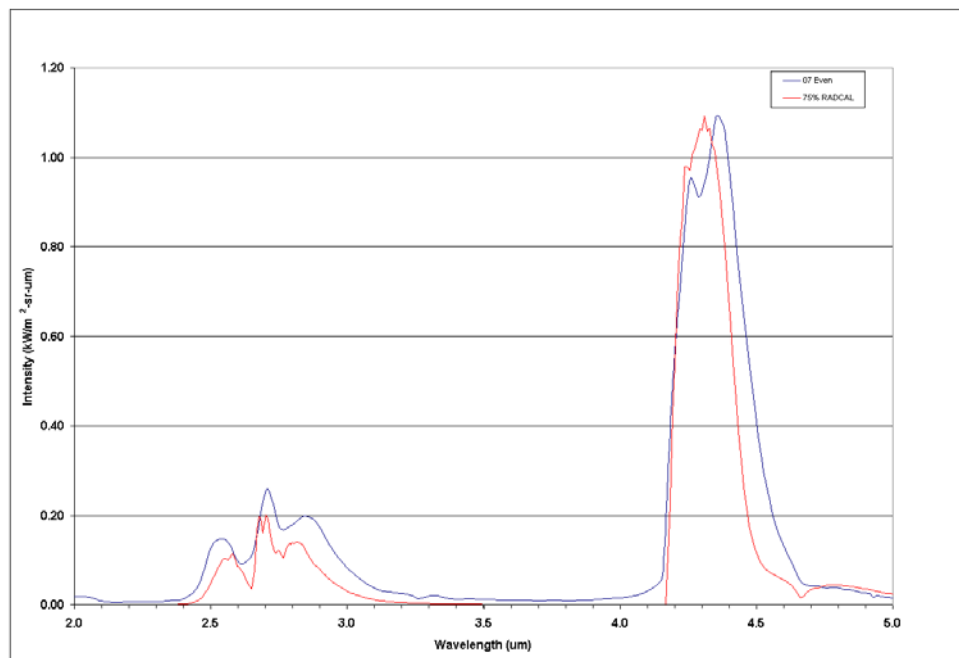
Figure 16: Comparison of Near-Centerline Emission Spectra with RADCAL Predictions. RADCAL Intensity Scaled-down 25%.

26

The resultant transmission scan spectra (I/Io) are plotted in Figure 17, with the attendant problems obvious for Scans 9 and 10. Transmissions greater than 1.0 are also apparent in all scans for the highest wavelengths, >4.7 μm. These trends with the transmission calculation are typical of the other data series. It is postulated that beam steering within the flame is causing the absorption beam to move slightly on the collection ball lens, dramatically changing the amount of energy that is coupled into the fiber. The sensitivity of coupling from the ball lens into the fiber was noted as a concern during alignment bench tests.

The near centerline transmission data is compared to RADCAL predictions, showing good agreement. The agreement could be improved if species concentrations were fit rather than assuming equilibrium concentration (the job of the tomography); however, the comparison shows that for the most part the absorption measurement produce proper trends. Unfortunately Series AD shows is by far the best agreement with the RADCAL predictions.

27
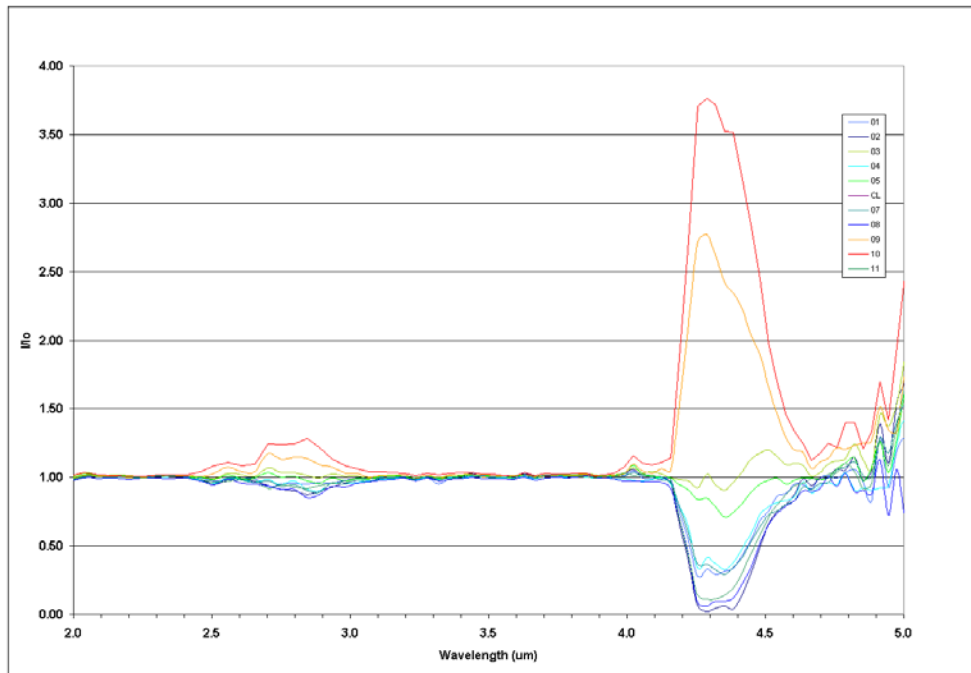
Figure 17: Transmission Data as a Function of Scan Position for Series AD Series.
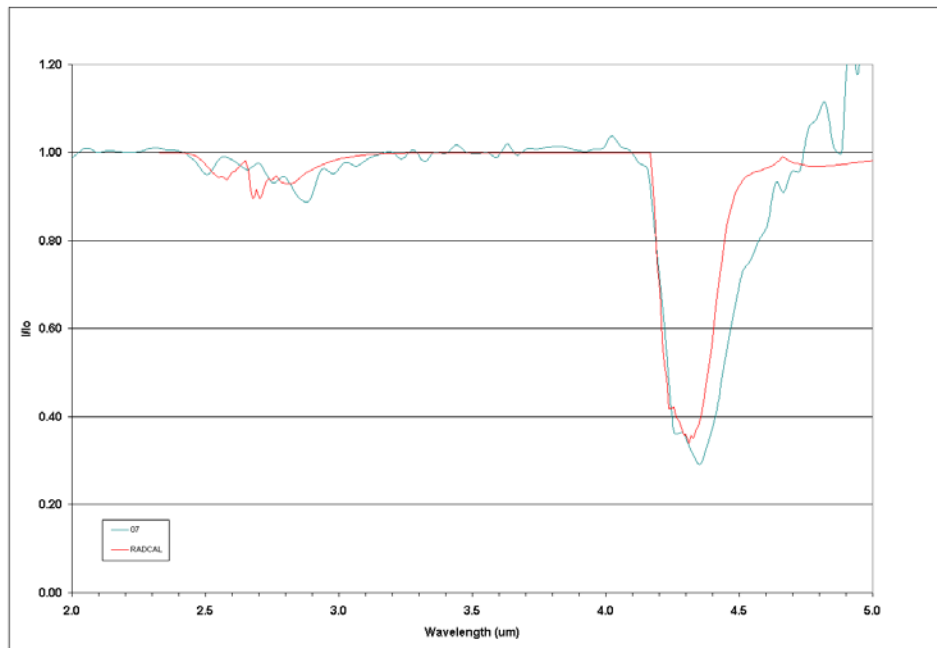
Figure 18: Comparison of Near-Centerline Transmission Spectra with RADCAL Predictions.

28
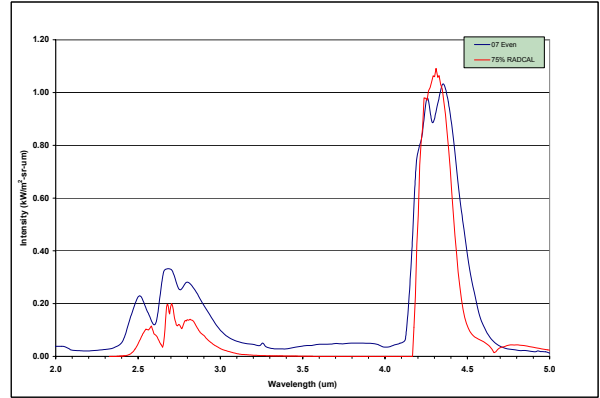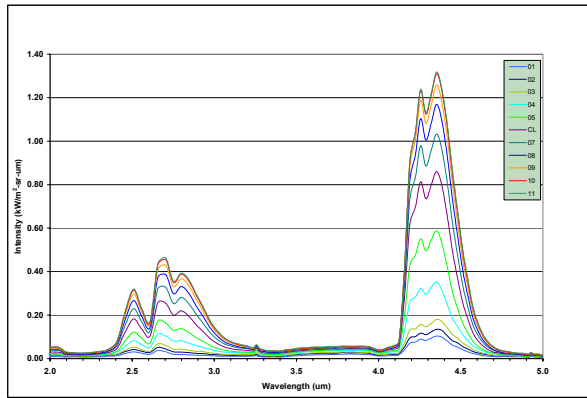
Figure 19: Even Pixel Emission Data (L) and Comparison of Near-Centerline Transmission Spectra with RADCAL Predictions (R) for Stoichiometric Condition (Series AA)
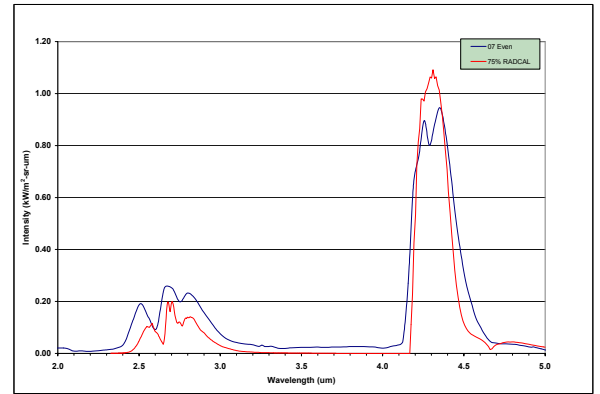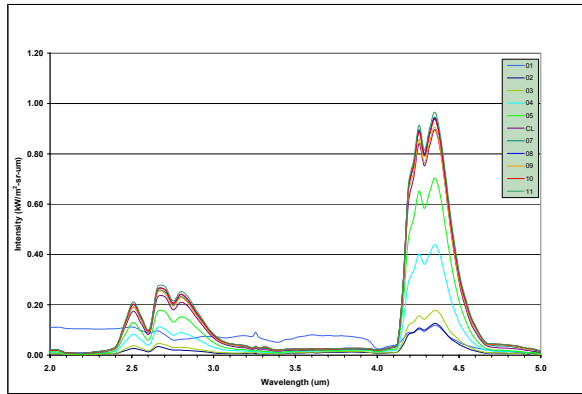
Figure 20: Even Pixel Emission Data (L) and Comparison of Near-Centerline Transmission Spectra with RADCAL Predictions (R) for Fuel-Rich Condition (Series AB)
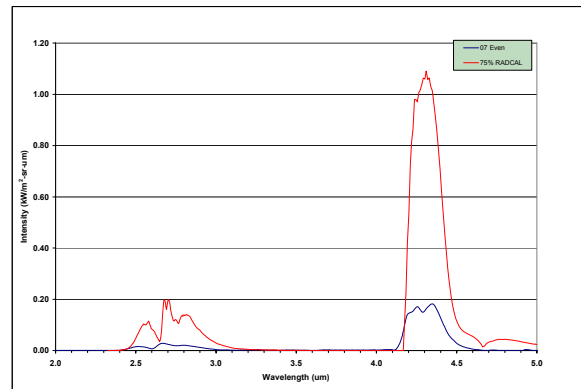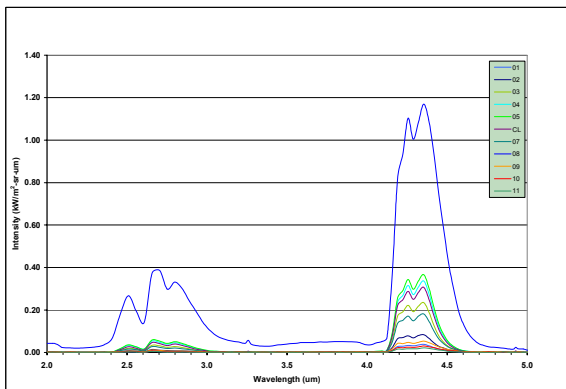
Figure 21: Even Pixel Emission Data (L) and Comparison of Near-Centerline Transmission Spectra with RADCAL Predictions (R) for Small Fuel-Lean Condition (Series AC)

29

180

Spectral Integration

The final data manipulation step required before the FAST program can perform tomographic reconstruction is integration of the emission and transmission data over significant wavelength bands. It is important to remember that each pixel integrates the incident radiation over a narrow wavelength range (20 nm).

The resultant measurement from the ES-100 spectrometer are an incremental irradiance, in units of kW/m^2-sr-μm. Multiplying the irradiance of a given pixel by the pixel detection bandwidth (20 nm) results in the total irradiance detected in the wavelength range covered by the specific pixel. Integrating irradiance (i.e., emission) over a wavelength range covered by several pixels is simply the summation of the individual pixel irradiance. Since only even pixel data are used during STIRS fiber optic data reduction, pixel data must be interpolated to estimate appropriate data for the wavelength ranges covered by the omitted odd pixels. The integrated emission intensity is reported for a wavelength range defined by the edge wavelengths for the extreme pixels used. For example, if the first pixel used in the integration is centered at 4.46 μm and the last pixel used in the integration is centered at 4.46 μm, the wavelength range covered by the set of 20 nm wide pixels is 4.44 to 4.48 μm. The resultant integration formula for emission N is:

$$N = \Delta\lambda(N_1 + N_n + 2\sum_{i=2}^{n-1} N_i)$$

(5)

Where N_i is the emission detected by pixel i, $\Delta\lambda$ is the pixel bandwidth (20 nm for the ES-100 spectrometer) and subscripts "1" and "n" are indexes for the first and last even pixels used in the integration.

The "integrated" transmission of the data plotted in Figure 23, Figure 25, Figure 27 and Figure 29 for the same wavelength range is actually an average of the individual pixel values. Again, the even pixel data are interpolated to estimate the transmission characteristics of the omitted odd pixel data, resulting in the following weighted average equation for band transmission τ:

$$\tau = 1.5\tau_1 + 1.5\tau_n + 2\sum_{i=2}^{n-1} \tau_i$$

(6)

Band data were developed for Series AA through Series AD covering each of the species associated with spectral features in the above emission and transmission plots for H_2O and CO_2. Referring to Figure 14, the left H_2O feature is integrated between wavelengths of 2.41 and 2.61 μm. The right CO_2 spectral feature covers 4.46 to 4.66 μm. Radial distributions (from burner edge to centerline) of the integrated emission and transmission are plotted for each of the bands in Figure 22 through Figure 29. As may be expected from the previous discussion, the integrated emission trends seem reasonable while the transmission data is at least suspect and likely useless.

30
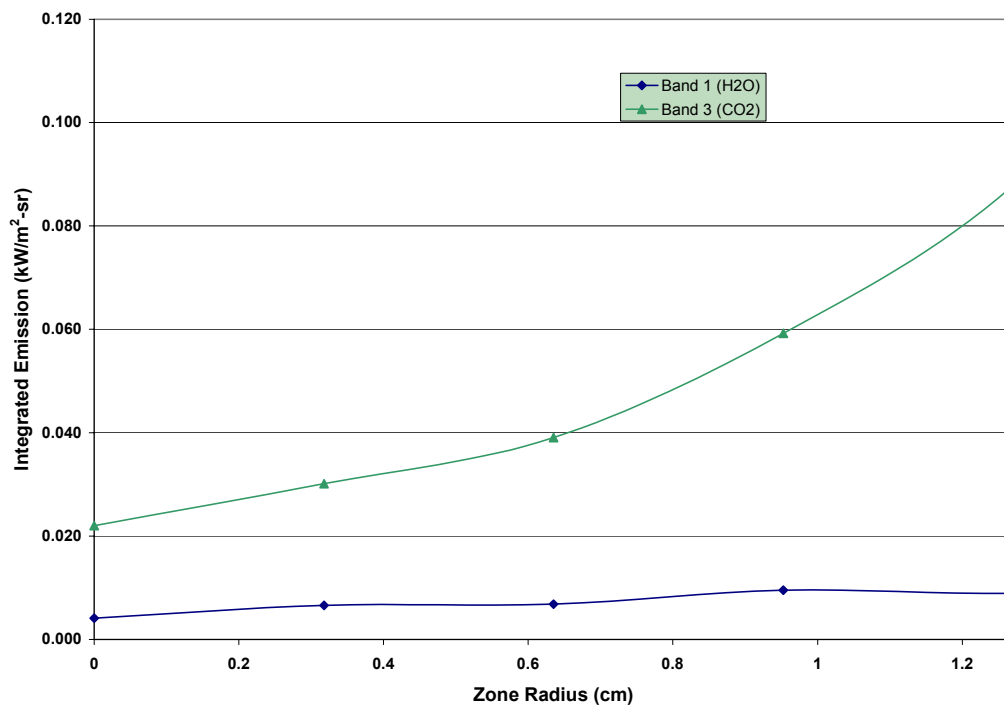
Figure 22: Integrated H_2O and CO_2 Emission for Stoichiometric Case (AA Series).
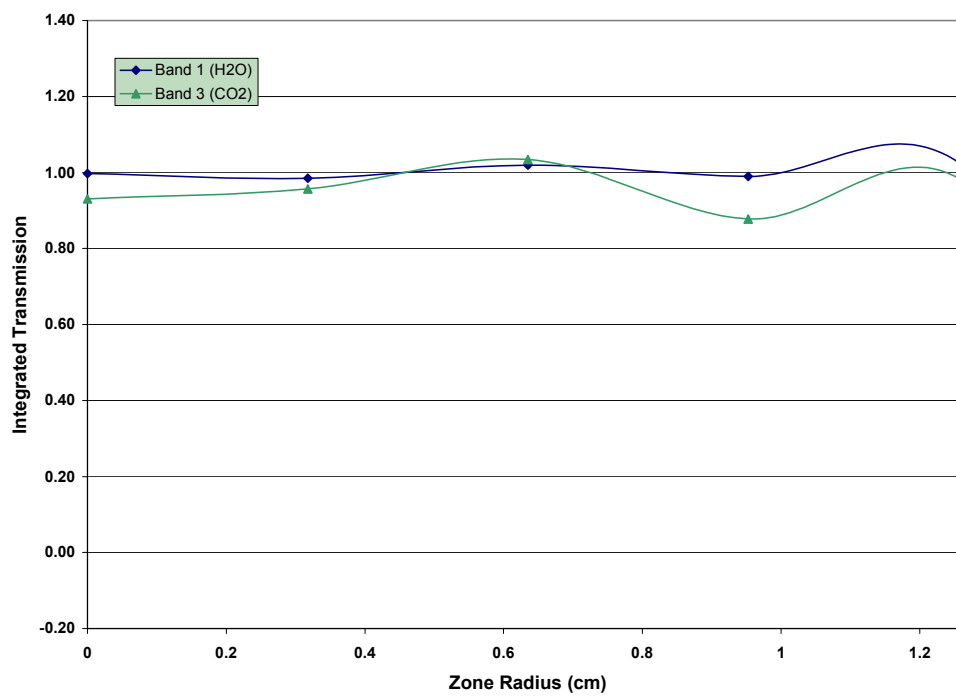
Figure 23: Integrated H_2O and CO_2 Transmission for Stoichiometric Case (AA Series).

31
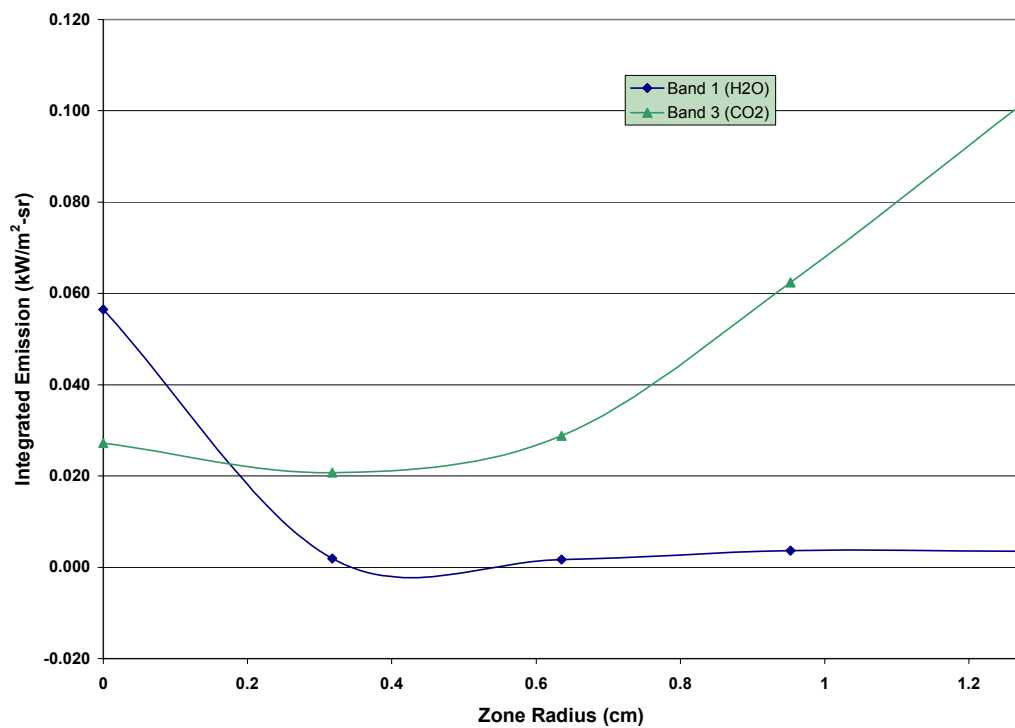
Figure 24: Integrated H$_2$O and CO$_2$ Emission for Fuel-Rich Case (AB Series)
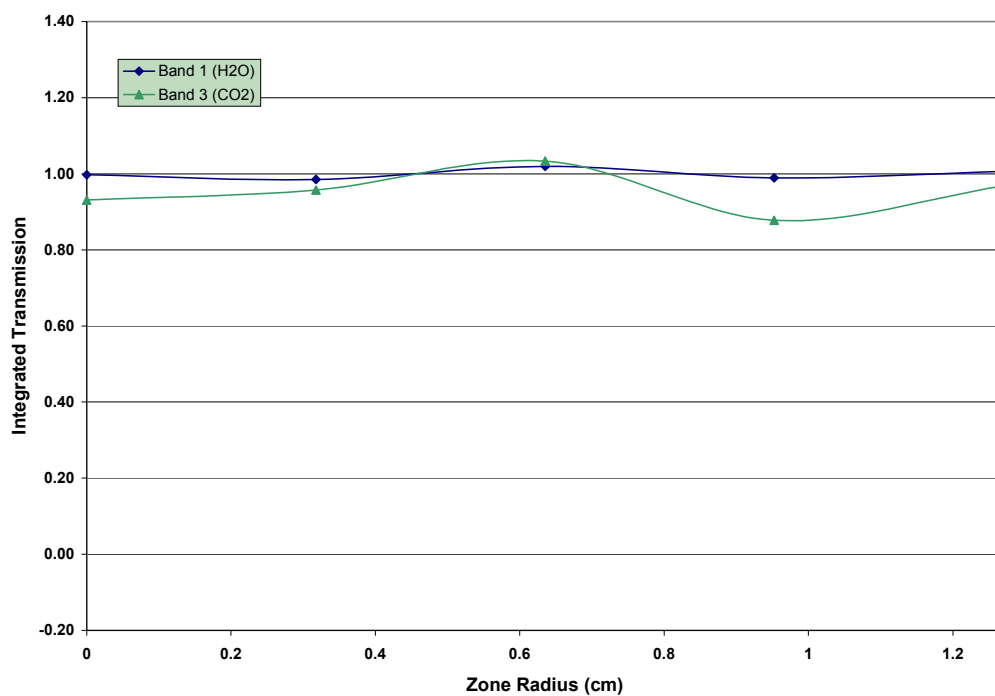
Figure 25 Integrated H$_2$O and CO$_2$ Transmission for Fuel-Rich Case (AB Series)

32
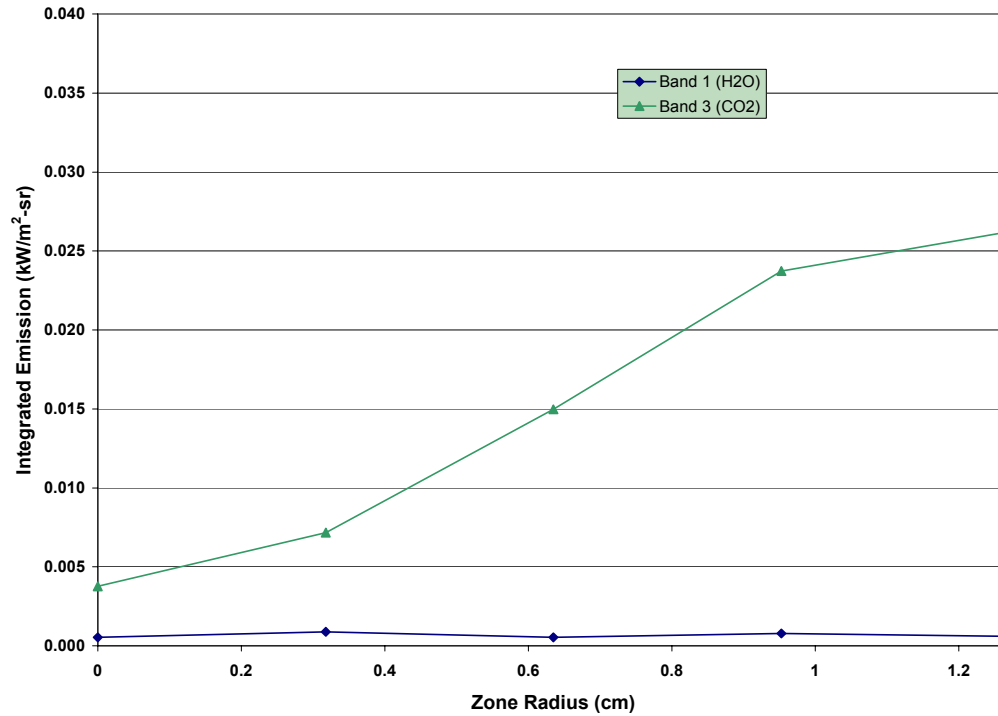
Figure 26: Integrated H$_2$O and CO$_2$ Emission for Small Diameter Fuel-Lean Case (AC Series).
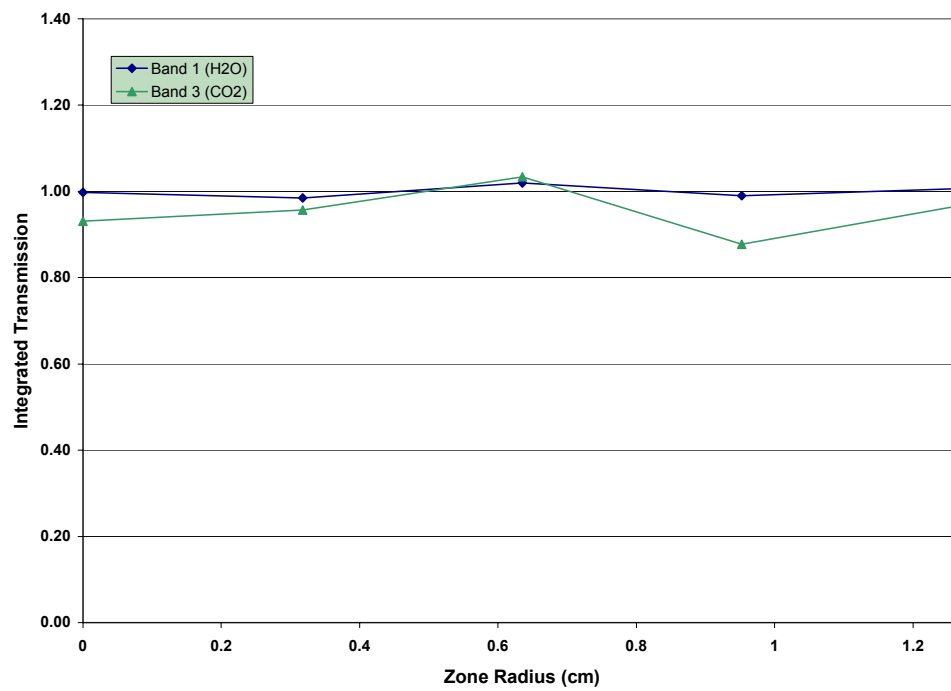
Figure 27: Integrated H$_2$O and CO$_2$ Transmission for Small Diameter Fuel-Lean Case (AC Series).

33
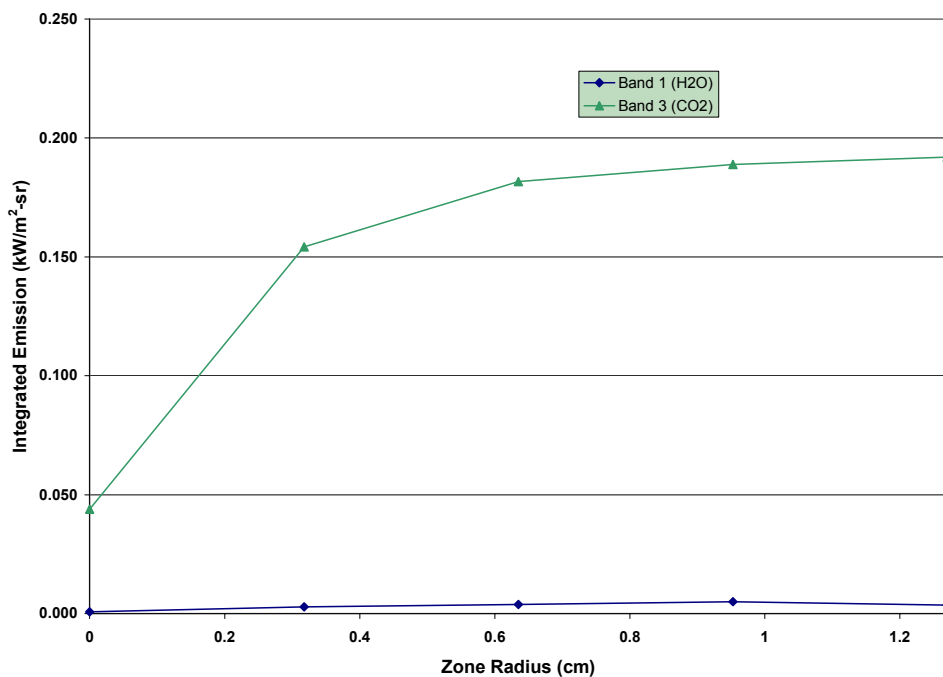
Figure 28: Integrated H_2O and CO_2 Emission for Small Diameter Fuel-Rich Case (AD Series).
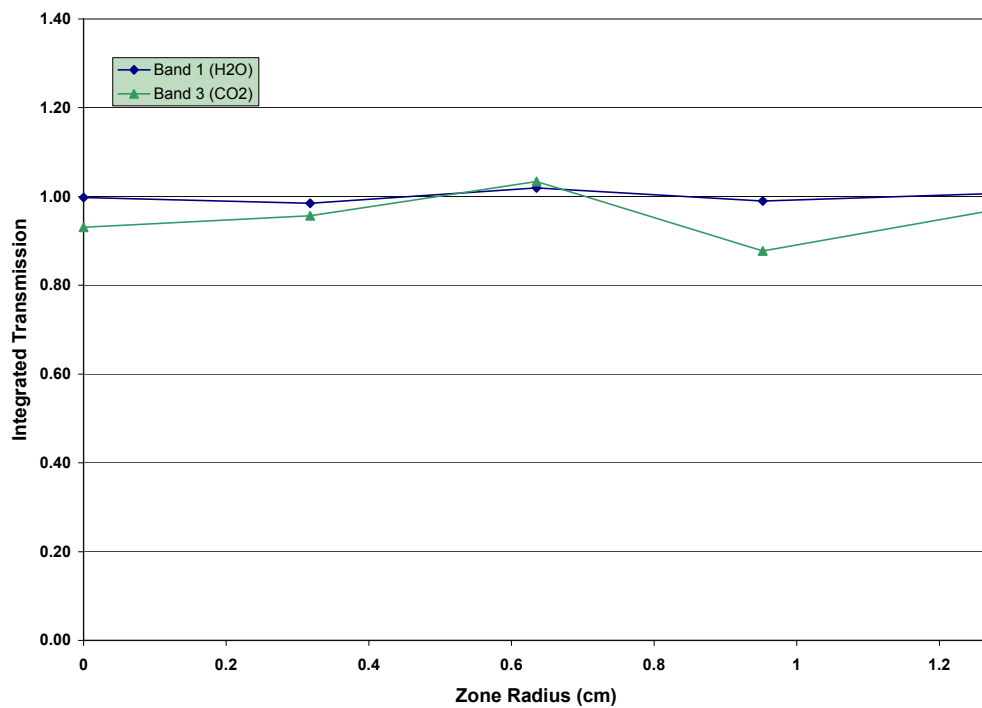
Figure 29: Integrated H_2O and CO_2 Transmission for Small Diameter Fuel-Rich Case (AD Series)

34

3.2.5 Tomographic Reconstruction

The final step is to perform the tomographic reconstruction of the zone species concentrations and temperature. The FAST program is used to reconstruct the flowfield using an onion-peeling deconvolution approach. The FAST program inputs include the integrated emission and transmission for each measurement station. Station definition data also includes local static pressure, assumed to be ambient at all stations for the example case. Additional inputs include estimates of temperature and species concentrations at each station. Equilibrium predictions were made using the one-dimensional equilibrium module of the TDK computer code[5] for equivalence ratios Φ of 0.8, 1.0 and 1.2 (Table 2). The estimated temperature drops to ambient outside the nozzle exit.

Table 2: Equilibrium Temperatures and Species Mole Fractions

Φ	Temperature (R)	CO_2	H_2O
0.8	3611	0.077	0.155
1.0	4021	0.086	0.185
1.2	3864	0.063	0.189

In an effort to simplify the reconstruction and data evaluation processes, wavelength bands were selected where interactions between CO_2 and H_2O are minimal, 4.46-4.66 μm for CO_2 and 2.41-2.61 μm for H_2O. The approach was to leave the emission data unaltered, if possible, and adjust the transmission to get a plausible, converged solution. The FAST code was not able to converge for all line of sight measurements using the initial integrated band data. Table 3 lists which data sets were adjusted to achieve convergence. Generally, the transmission had to be adjusted down for each species and series, except for H_2O transmission in the AA series, which converged without any changes. This is not surprising since, as discussed above, the transmission values calculated from the measurements often exceed 1.0.

Figure 30 through Figure 37 compare deconvolved temperature and species mole fractions for H_2O and CO_2 for each series. The radius increases from the burner edge (0 cm) to the centerline (1.27cm). A horizontal dashed line is included for the equilibrium concentration of each species to facilitate comparison. In general the trends are reasonable. However, the water prediction for Series AB and AC have problems. Series AB was the worst data series and required the most manipulation and changes in emission and transmission. However, the reduced AB series had the best match of CO_2 concentration as compared to equilibrium calculations.

Figure 33 and Figure 36 show a mole fraction greater than 1.0 for H_2O at a radius 0.635 cm is show on both plots. This is obviously an incorrect result, but it was an output from the FAST program. The code obviously needs to be corrected to prevent this solution from occuring.

Table 3: Data Manipulations Required for Convergence of Tomographic Reconstruction

Series	Species	Emission 1	2	3	4	5	Transmission 1	2	3	4	5	Φ	Diam.
AA	H2O	-	-	-	-	-	-	-	-	-	-	1.0	1.9
AA	CO2	-	-	-	-	-	√	√	√	√	√	1.0	1.9
AB	H2O	√	√	√	-	-	-	√	√	√	√	1.2	1.9
AB	CO2	√	-	-	-	-	√	-	-	-	-	1.2	1.9
AC	H2O	-	-	-	-	√	√	√	√	√	√	0.8	1.125
AC	CO2	-	-	-	-	-	-	√	√	-	√	0.8	1.125
AD	H2O	-	-	-	-	-	√	√	√	√	√	1.2	1.125
AD	CO2	-	-	-	-	-	√	√	√	√	√	1.2	1.125

√ = Changed

-- = Used as measured
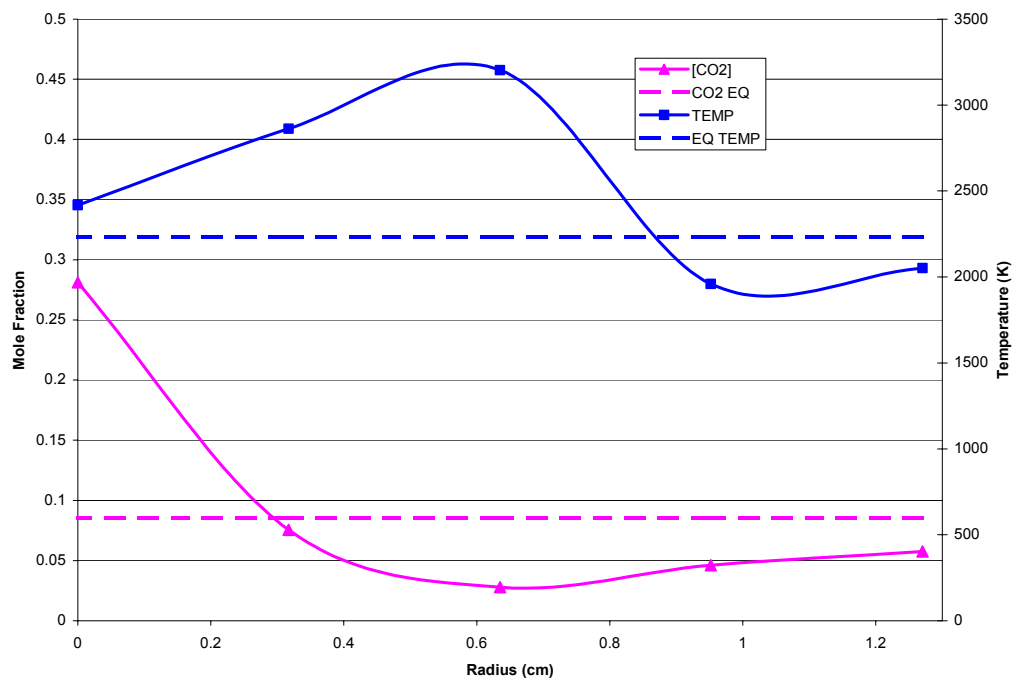
Figure 30: Comparison of Equilibrium Zone Temperatures and CO_2 Concentration with FAST Predictions (AA Series)
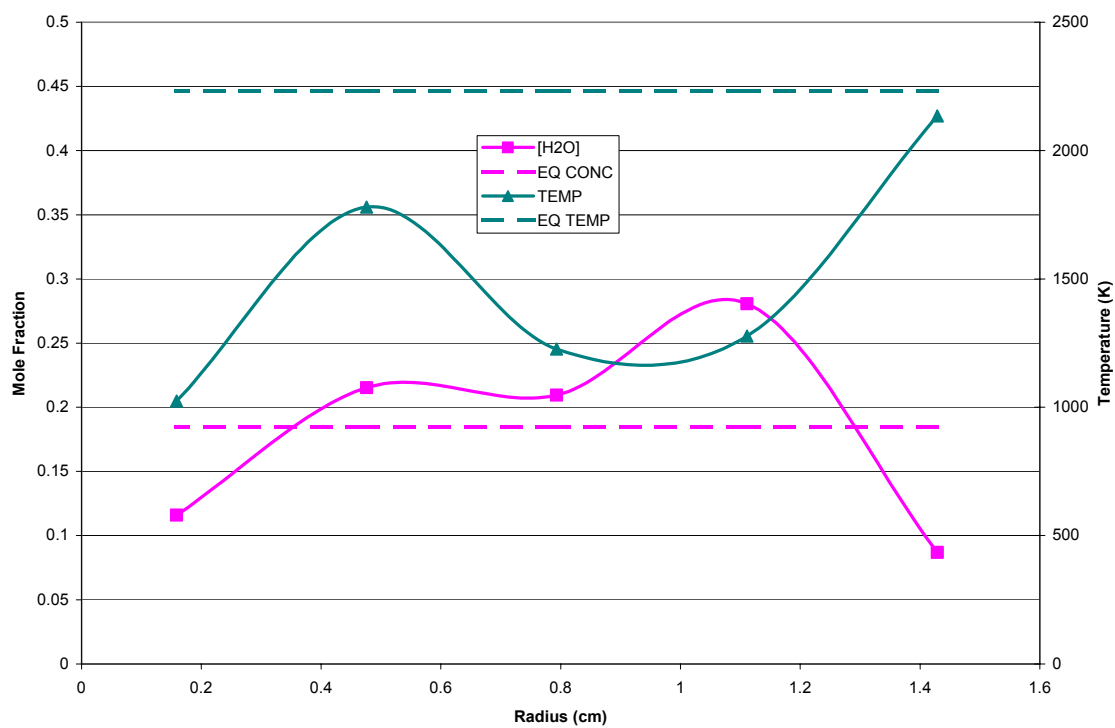
Figure 31: Comparison of Equilibrium Zone Temperatures and H_2O Concentration with FAST Predictions (AA Series)

37
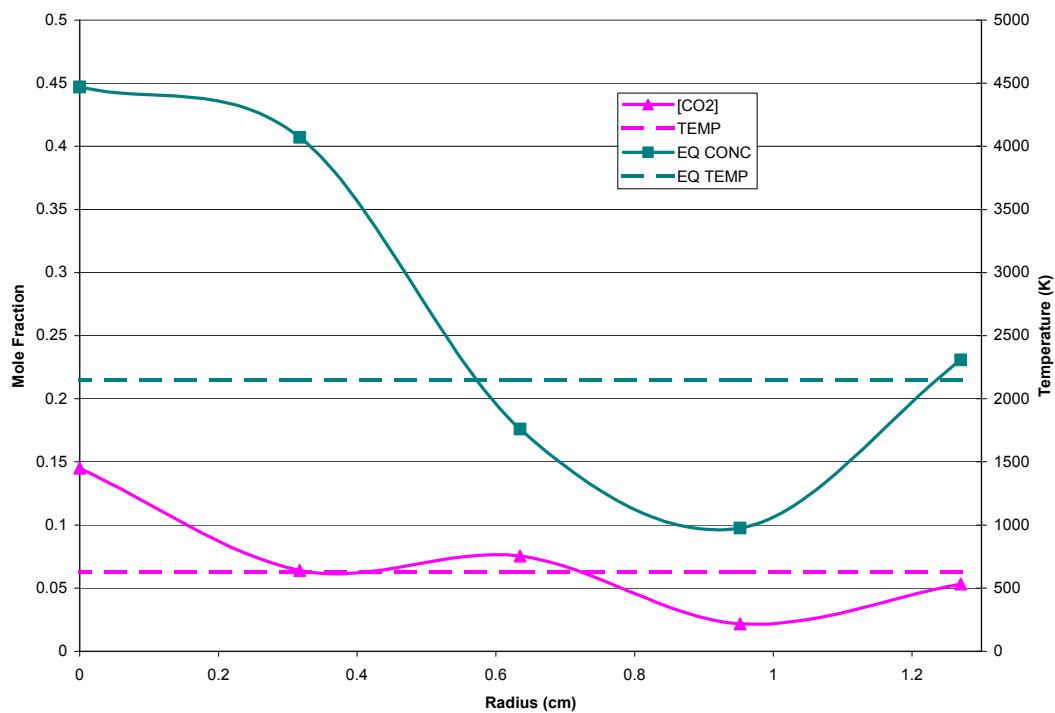
Figure 32: Comparison of Equilibrium Zone Temperatures and CO$_2$ Concentration with FAST Predictions (AB Series)
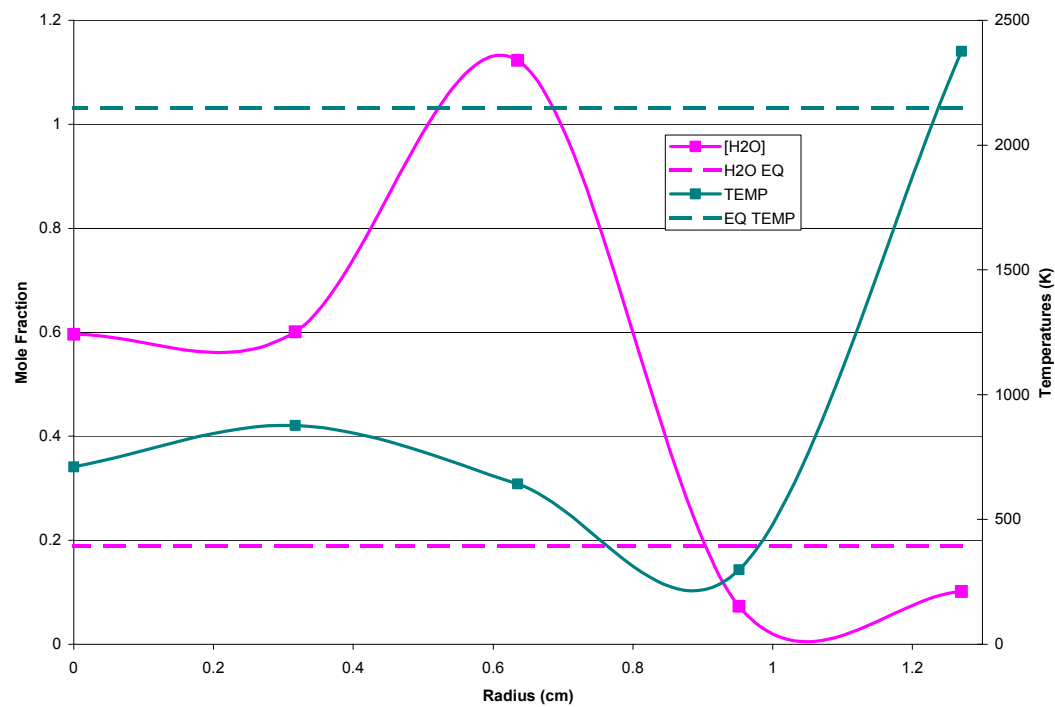
Figure 33: Comparison of Equilibrium Zone Temperatures and H$_2$O Concentration with FAST Predictions (AB Series)

38
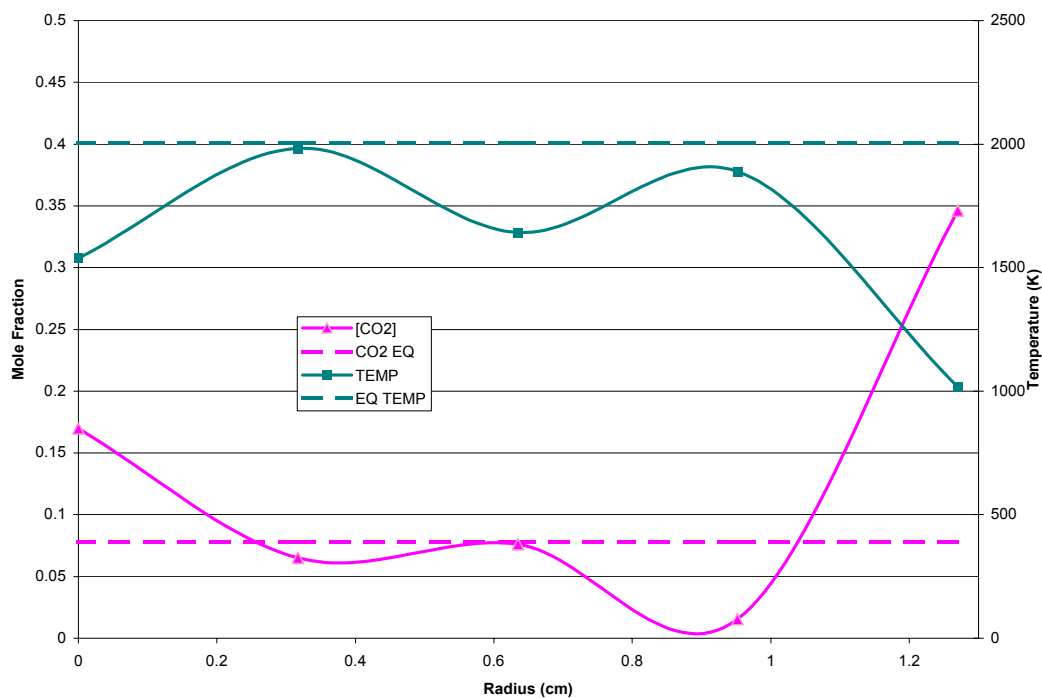
Figure 34 Comparison of Equilibrium Zone Temperatures and CO_2 Concentration with FAST Predictions (AC Series)
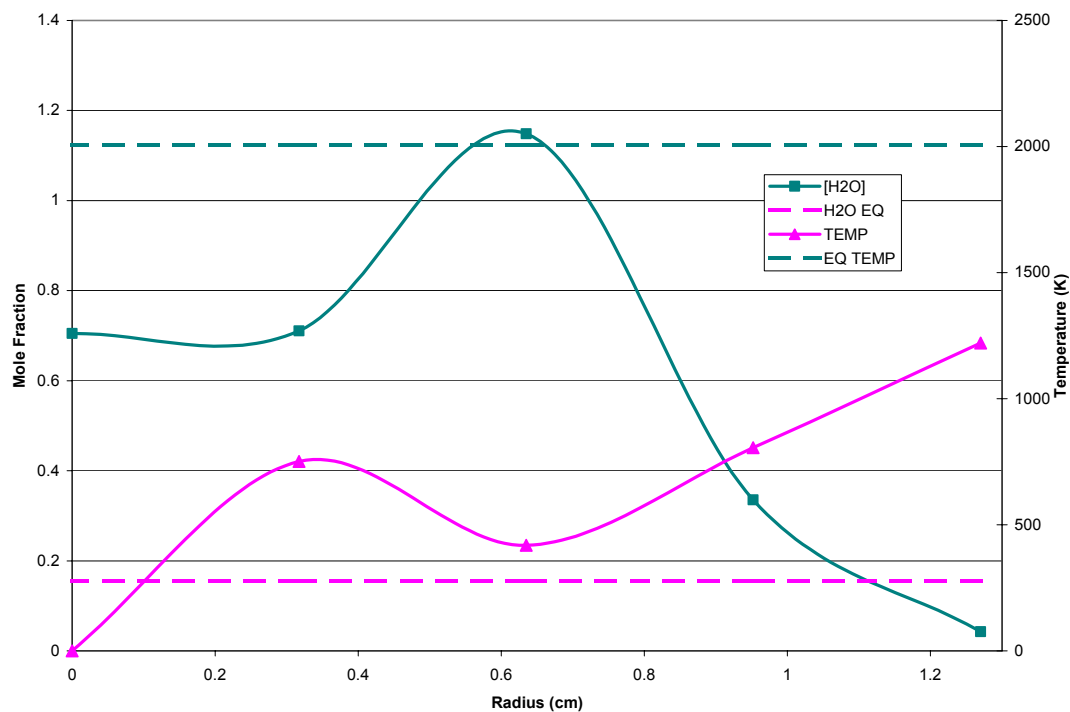
Figure 35: Comparison of Equilibrium Zone Temperatures and H_2O Concentration with FAST Predictions (AC Series)

39
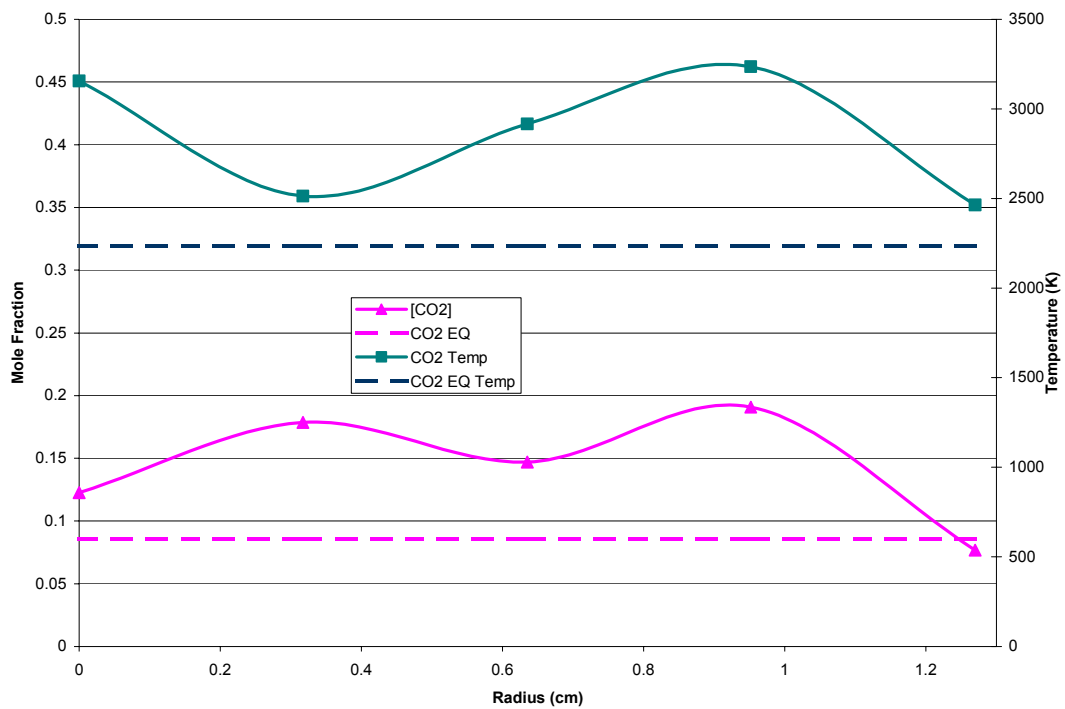
Figure 36: Comparison of Equilibrium Zone Temperatures and CO_2 Concentration with FAST Predictions (AD Series)
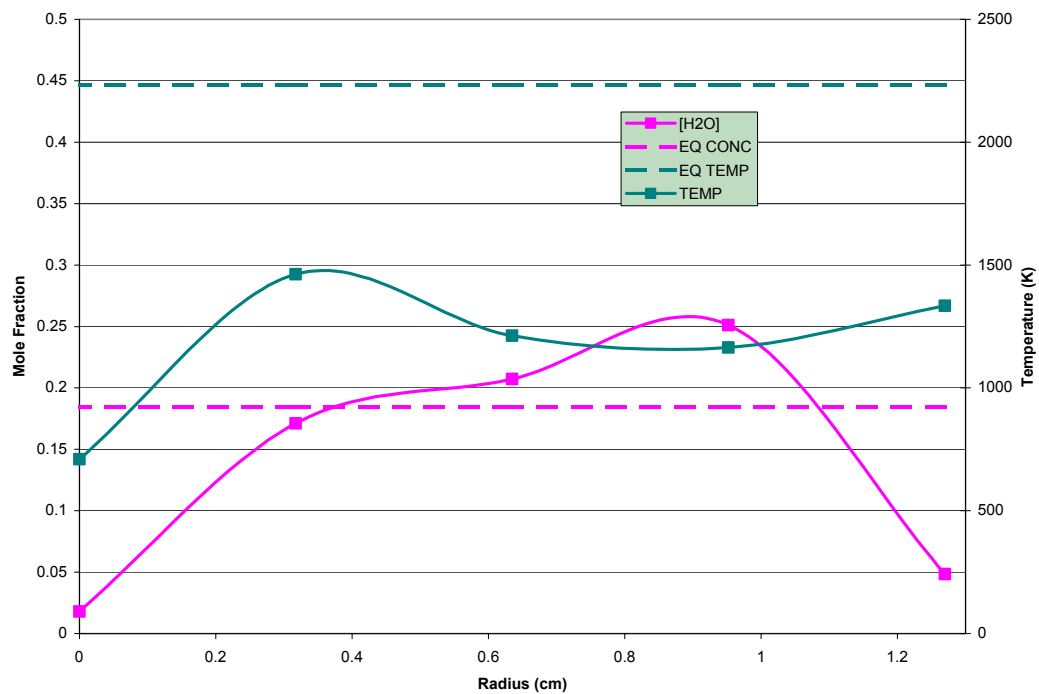
Figure 37: Comparison of Equilibrium Zone Temperatures and H_2O Concentration with FAST Predictions (AD Series)

40

4.0 REFERENCES

[1] www.spectraline.com

[2] www.reflexusa.com

[3] Charschan, S.S.; Lasers in Industry, Western Electric Series, (Western Electric Company), Van Nostrand Reinhold Company, New York, N.Y.,1972

[4] Muss, J.A. and Higgins, B.S.; "A Scanning Infrared Emission-Absorption Diagnostic for Measurement of Rocket Exhaust Plume Temperature and Species Distribution", NASA ASDC Propulsion Measurement Sensor Development Workshop, Huntsville, Alabama on May 13-15, 2003

[5] Nickerson, G.R., Dunn, S.S., Coats, D.E. and Berker, D.R.; *Two Dimensional Kinetics (TDK) Nozzle Performance Computer Program User's Manual*, SN-168, Software and Engineering Associates, Carson City, NV, January, 1999

Correlation of Milk Composition and Fouling with Biofilm Formation and Microbial Spore Production in Heat Exchangers

Rafael Jiménez-Flores
Dairy Products Technology Center

Yarrow Nelson
Civil and Environmental Engineering

Dan Walsh
College of Engineering

Cal Poly State University

California Central Coast Research Partnership

Report

Title of proposal   **Correlation of milk composition and fouling with biofilm formation and microbial spore production in heat exchangers.**

Investigator(s) and Department(s)

Rafael Jiménez-Flores	Dairy Products Technology Center
Yarrow Nelson	Engineering
Dan Walsh	Engineering

Summary of Achievements:

- Two systems were built in where to study fouling and biofilm deposition. One at the laboratory scale, and one in the DPTC pilot plant.
- The pilot plant system was tested in 20 to 25 litter batches, and for extended runs of up to 10 hours.
- The laboratory scale system has a volume capacity of 5 liters, but can be run for extended periods of up to 24 hours.
- All the DPTC dairy cows were classified by PCR-based techniques for their genetic variant k-casein and β-lactoglobulin. Milks from these cows was selected and used for these experiments.
- Pilot plant system's parameters were optimized for measurements of heat transfer fluctuations, and heat transfer efficiency. The basis of these measurements is accurate and continuous measurement of inlet and outlet temperatures from the plate heat exchanger.
- At the pilot plant scale we could determine that there is a fundamental difference on the nature of film deposition of the different genetically classified milks. Tendencies agreed with reported values.
- At the laboratory scale it was proven that the rate of fouling and fundamental nature of biofilm formation depends to an important extent on the material considered. It appears that material deposition from milk at the surfaces precedes bacterial deposition, and point proliferation.
- Microscopically, the laboratory system allowed observations of biofilm and bacteria attachment, and yield useful information on the interaction of milk and bacteria as a function of time and material employed.
- Two master's students were able to finish their thesis and research work, and the involvement of at least 5 undergraduate students helped in this project.

Expanded Report

       Recent research by the New Zealand Dairy Board suggested that fouling during milk processing could be reduced by using classified genetic variant phenotype cows that produce specific variants of β-lactoglobulin (BLG) and κ-casein. Because of the important role of biofouling in increasing the operating costs of milk processing and possible public health issues, the effect of genetic variants on biofouling was further investigated in a joint study between the College of Agriculture and the College of Engineering at Cal Poly. A pilot-scale heat exchanger was assembled and used for measuring biofouling rates for different types of milk from genetically classified Cal Poly dairy cows. This apparatus was used to determine biofouling rates by monitoring both milk and hot water inlets and outlets temperatures using thermocouples connected to a data logger. Biofouling was determined based on the changes in delta T (inlet hot water and milk outlet temperature difference), milk outlet temperature, and heat transfer rate. Biofouling rate was also analyzed in terms of key components in the biofilm such as protein, mineral, and fat as well as total dry weight. These measurements were measured at the pilot plant scale, and a system was built and optimized at the DPTC. We have reported on this previously.

       Another aspect of interest in this work was the analysis of the interactions of the different genetic variant proteins in milk and the surface of different materials used for the construction of plate heat exchangers. For this objective, a laboratory system was designed and used with different samples of materials with practical and potential use in the food industry.

   The specific objectives of this project included:
1. Assemble a pilot heat exchanger set up that would measure milk biofouling rate.
2. Identify the effect of milk-type genetic variations (BLG AB, BLG BB and BLG AA) on the formation of biofilm and to measure and compare their biofouling rate in the plate heat exchanger.
3. Characterize the biofilm formed by fouling of plate by a flow cell model.
4. Test different types of metals/alloys in the laboratory flow cell model to see any changes of biofilm production.
5. Find any changes in surface-active properties of the milk by measuring any differences in the contact angles between milk and a common polished surfaces (stainless steel 316. 304, 430 and Titanium).
6. Develop methods to analyze the composition of material (protein, fat, and
7. mineral) deposited during heat processing of milk-type genetic variants.
8. Employ statistical methods to analyze results objectively.

MATERIALS AND METHODS

<u>Pilot-scale heat exchangers</u>

Biofouling experiments were carried out on a pilot plant milk pasteurizer. Initially a portable plate heat exchanger was used for the fouling experiments, but this model system was inadequate because of difficulty keep controlling the inlet hot water temperature. Also, this system sent the tested milk into the drain, and consequently required large volumes of raw milk to be tested and created waste disposal issues. Therefore, a new system was constructed the heat exchanger from an existing pasteurizer system (called PMS). Thermocouples and a positive displacement pump were attached to this system. The advantages of the new system were: (1) hot water temperature was regulated by an automatic control system, (2) less amount of milk was drained to sewage (about 15 gallons at the end of each run compared to 60 gallons) by recirculating the milk.

The equipment used was a heater- cooler pilot plant- HTST (high temperature, short time) PMS (from Processing Machinery & Supply Co., Philadelphia). The heat exchanger consisted of stainless steel plate heat exchangers (Junior from APV Crepaco, Inc.) made up of two exchange sections (heating and cooling) (Figure 4.1), which each consist of a number of parallel flow channels formed by metal plates, which are separated by gasket material around the perimeter of each plate. Nozzles for the flow of fluids extend through the frames to the plate packages. Two storage tanks for chilled and hot water were used to maintain constant temperatures. Heating and cooling units were used for production of chiled and hot water. A control console included a hot water temperature regulator. This apparatus was used to determine biofouling rates for different types of milk products by monitoring the increase in the temperature differential as indicated by lowering milk outlet temperatures and also an increase in delta T (temperature difference between hot water inlet minus milk outlet temperature). The apparatus design was based on pilot scale milk pasteurization unit at Cal Poly's Dairy Product Technology Center (DPTC) currently available for research applications.

To measure temperature differentials, four thermocouple probes (NPT series type K) were installed onto inlets and outlets. A data logger (OM-3001 from Omega Engineering) was used to record all four temperatures (two inlet plus two outlet).

Figure 1 Pilot scale milk pasteurizer at Dairy Product Technology Center (DPTC)

Figure 1 shows the process flow diagram for the experimental setup. The product (milk) at about 40 ˚F and a flow rate of 0.5 GPM is supplied from a product tank and then pumped to the heating section through the plates and then recirculated to the product tank. Hot water on the other side of the plates at about 206 ˚F and 15 GPM is pumped from the heating medium tank through the plates and then is recirculated to the tank where it is maintained at constant temperature using steam.

Figure 2 Schematic representation of the mix heater and cooler

Figure 2 shows the flow pattern and configuration in heating and cooling sections of plate heat exchanger. Nineteen plates along with four passes were arranged in the heating section to achieve operating conditions as given in Table 1.

Table 1 Processing conditions in the mix heater- cooler pilot plant- HTST from Processing Machinery & Supply Co. (PMS)

Number	Stream	Flow rate (GPM)	Temperature (F)
1	Milk inlet	0.5	40
2	Milk outlet	0.5	201
3	Hot water inlet	10	203
4	Hot water outlet	10	198
5	Cooling water inlet	10	39
6	Cooling water outlet	10	---
7	Product (milk) outlet	0.5	42
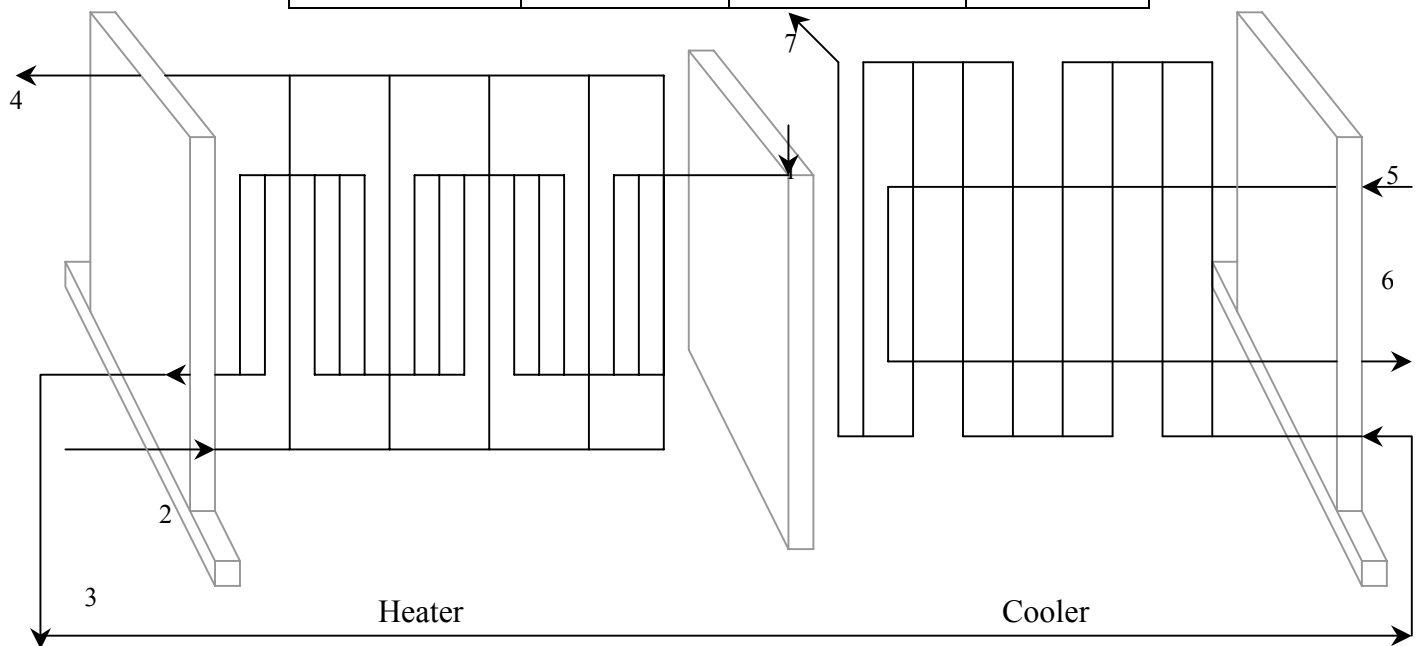
Figure 3 Configuration of the two heating and cooling units of the exchangers

The heating section of the heat exchanger was particularly prone to fouling. It was composed of nineteen plates with a total exchange area of 0.03 m^2 for each plate as shown in Figure 4.4.
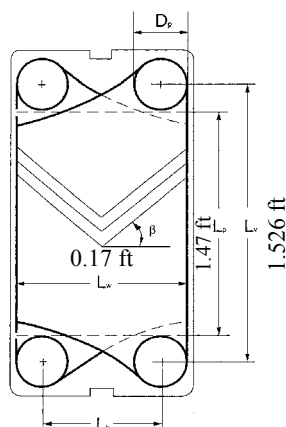
Figure 4 Main dimension of a chevron plate heat exchanger

Milk types used in biofouling experiments

Three different types of milk (control AB, BLG BB, and BLG AA) were obtained from Cal Poly dairy cows for this study. Milk types classification were as follows: (1) Control AB contains a mixture of κ-casein genetic variants (AA and BB) and a mixture of BLG genetic variants (AA and BB). (2) BLG BB contains a mixture of κ-casein genetic variants (AA and BB) and only BLG BB (3) BLG AA contains a mixture of κ-casein genetic variants (AA and BB) and only BLG AA. The characteristics of these milk types, as measured immediately before experimentation, are given in Table 4.2.

Table 3 Data indicating quality of milk used and processing conditions

Characteristic	Control AB			BLG BB			BLG AA		
	1	2	3	1	2	3	1	2	3
PH	6.86	6.61	6.71	6.64	6.64	6.71	6.58	6.75	6.66
fat content %	4.2	4.2	4.2	4.5	4.8	4.6	4.6	4	4.5
Total solid % (w/w)	13.08	12.97	13.20	13.30	13.83	13.65	13.42	12.81	13.79
Processing time (h)	6 ½	6 ½	6 ½	6 ½	6 ½	6 ½	6 ½	6 ½	6 ½
Milk flow rate GPM	0.51	0.48	0.49	0.52	0.51	0.51	0.48	0.51	0.51
Avg. Milk inlet temperature in heating section (F)	42.16	42.31	43.46	41.93	47.57	42.47	42.06	43.10	42.48
Avg. Hot water inlet temperature in heating section (F)	203.80	203.63	204.35	204.35	203.21	203.46	203.94	204.15	203.96

Laboratory Scale System

The working model for biofilm formation was machined from an aluminum block. Surplus hardware was used to manufacture the fittings to and from the model. The design of the small heat exchanger was a housing made of one piece of aluminum machined so that a smaller plate can be fitted into the middle of the hollowed section. Cover plates for the water and milk sides were manufactured from single sheet of rubber gasket and a carbon fiber plates added for rigidity. All of the hardware of the heat exchanger was fitted with stainless steel 440 screws. Poly tetra-fluoro- ethylene (PTFE) was used in the chamber to isolate the plate from the aluminum frame. Figure 5 shows the flow cell in a new and clean condition.

Figure 5.This is the miniature heat exchanger that was custom made for the project. The view is directly into the milk side showing the sample plate in place. The sample plate loaded is that of a thick stainless steel 316, and only required 4 screws to make the seal between the water and the milk. Later two more screws were added to stop leaks that were formed due to flexing on the thin plates.

The completed unit assembled for laboratory experiments is depicted in Figure 6. This set up allows experiments to run for 12 to 24 hours; temperature range from ambient to 93 C. The tubing and pumps were fitted to assure continuous run, and the system can be expanded to run several chambers at once.

Flow through the system can be controlled accurately from 5 to 500 ml/min; pH monitoring and temperature is continuous. Temperature controllers allow a stable heating with thermal plates, and due to the volumes of milk and water, this is achieved within one degree of fluctuation once the system is in a steady state.
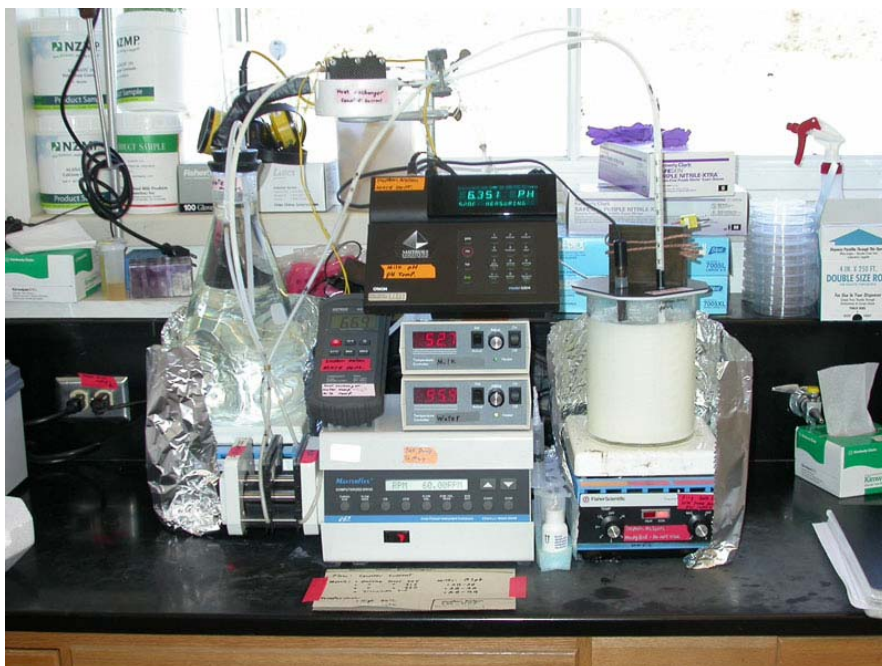
Figure 7.This is an image of the heat exchanger system used to form fouling of metal plates with raw milk. Shown is the two temperature controlled tanks one for milk and one for the heating water and their controllers on top of the peristaltic pump. The small meter to the left of the temperature controllers is the thermocouple meter to monitor the temperature inside of the heat exchanger. Note the heat exchanger is very close to the water flask to prevent any heat loss.

In a related series of experiments, the surface activity of the different milks in relation to the materials used in the heat exchangers will be measured. The contact angle between a droplet of fluid and the surface is a direct function of the surface activity interaction between the material containing the droplet and the droplet substance(s).
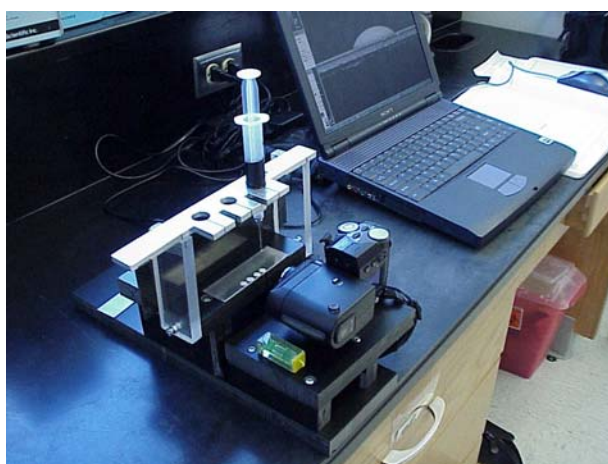
Figure 10. The complete system for measuring the contact angle of the milk. The milk is loaded into the syringe and placed into the holder. Gentle pressure is put onto the plunger and a single drop is released. Five seconds is let pass and then the digital camera takes an image of the side profile. The image is then loaded on the computer and an image analysis program is used to measure the contact angle.

Measurement of biofouling

Temperature monitoring

The rate of fouling on the heat exchanger surfaces was determined by monitoring the rise in temperature difference (Delta T) between the outlet milk and the inlet hot water, and also the decrease in milk outlet temperature, after 6 ½ hours run time. Temperature measurements were made every 30 seconds by use of the data logger.

Three replicate experiments were performed for each of the three milk types as described above in randomized order. Consequently there were nine runs in the heat exchanger. Multiple regression method (under Minitab software) was employed to analyze any statistical differences between milk types

All foulants were analyzed for protein, fat, moisture and mineral content using routine procedures at Cal Poly DPTC as described below. After each run the PHE was stripped and the plates and pumps were cleaned with caustic (Principal, Ecolab, MN), acid (HD Acid PL-10, Ecolab, MN), and cleaning in place solution (CIP acid sanitizer, Mandate, Ecolab, MN). Deposits were collected by scrapping off half the area of the second plate in the heating section of the heat exchanger, the place that has the highest temperature difference between hot and cold fluids (cold milk and hot water).

Biofilm analyses

Biofilm deposition rates were determined gravimetrically by scraping off half the area of the first plate heat exchanger (from top to the middle) in the heating section. These solids were dried at 100 ˚C for 4 hours in a vacuum oven and then weighed. In order to develop a better understanding of the mechanisms of formation of milk biofilms, proteins, minerals, and fat deposited during biofoulling in pilot-scale system were analyzed. For these analyses, wet attached deposits were scraped off of the plate heat exchanger surface that had direct contact with raw milk while processing.

Quantitative protein analyses

The Kjeldhal method was used to determine the nitrogen content of the milk deposits, and the percent protein was calculated from the nitrogen content (AOAC 1995 c). 0.5 g of the dried biofilm (collected in the previous section) was used for each protein analysis. Samples were digested in 20 mL H_2SO_4 using 3 tablets of LCT-40 Kjeldhal (Fisher Chemicals, NJ, USA) catalyst to each digestion flask. Each flask was heated by an electric burner from low to high setting (190 °C for 45 min, 300 °C for 45 min, and 425 °C for 75 min) with a fume ejection system on. This digestion releases nitrogen from the proteins and produces ammonium salt. NaOH is then added to hydrolyze ammonium and release NH_3, which is distilled, collected in H_3BO_3 solution and titrated with 0.1 N HCl to a pink endpoint (Figure 4.5).

<div align="center">

DIGESTION   **DISTILLATION**

</div>

<div align="center">
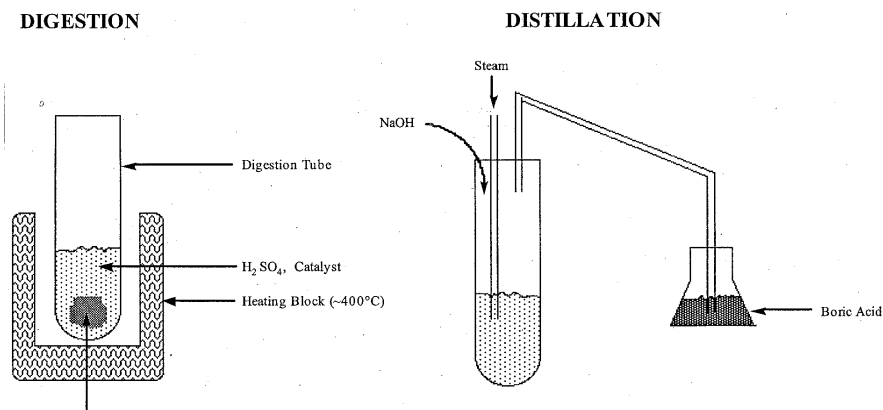
Figure 4.5 Digestion and distillation in Kjeldhal method

</div>

Protein content was calculated by assuming that milk protein is 15.7 percent nitrogen, which is standard for milk proteins (Jones 1931). Protein content for each sample was calculated as follows:
(Assumption for the titration, X mL of 0.1 N HCl was required).

$$\% \ (w/w) \ N = \frac{X \ mL \ HCl}{sample \ weight \ (gr)} \times \frac{0.0001 \ mole \ HCl}{mL \ HCl} \times \frac{1 \ mole \ NH_3}{1 \ mole \ HCl} \times \frac{14 \ g \ N}{1 \ mole \ NH_3} \times 100$$

$$\% \ Protein \ = \% \ Nitrogen \ \times 6.38$$

Qualitative protein analyses

Milk proteins deposited during biofouling were analyzed qualitatively using electrophoresis with SDS-PAGE (Sodium Dodecyl Sulfate-Poly Acrylamide Gel Electrophoresis). This method separates, compares, and characterizes proteins based primarily on their molecular weights (Laemmli 1970). SDS is a detergent that can dissolve hydrophobic molecules but has also a negatively charged sulfate group attached to it. SDS binds to hydrophobic portions of a protein, disrupting its folded structure and allowing it to exist stably in solution in an extended conformation. As a result, the length of the SDS-protein complex is proportional to its molecular weight. In SDS-PAGE, for electrophoresis analysis, a slab of polyacrylamide is placed in a solution of the proteins to be analyzed and an electrical charge is generated across the gel. Protein will migrate through this gel at different rates depending on their charge to mass (length) ratio. After a specified time for migration, the proteins in the gel are stained, allowing visualization of bands associated with each protein. By comparing these bands with standards, the protein can be identified (Bollag *et al.* 1996).

In reducing SDS-PAGE a reducing agent such as β-mercaptoethanol is added break up all polypeptide disulphide bonds. Consequently, in non-reducing condition covalent polypeptide bonds in protein remain intact (Scopes 1982).

The following procedures were developed for protein electrophoresis of the biofilm samples: (1) Resolving (separating) gels were prepared with 12 % Acrylamide. (2) Stacking gels were prepared with 4 % Acrylamide. (3) The biofilm scrapings were

diluted to 10 % total solid by dissolving 1 g in 10 mL DI water, and then grinding with a Ultra-Turrax T8 (IKA-WERKE,GMBH & Co. kg, Germany) for better dissolving of solids. (4) Protein samples were adjusted/normalized to approximately 0.5 mg/mL both for whole milk and biofilms sample. (5) One part of the protein samples were added to three parts Laemmli sample buffer (both reducing and non-reducing conditions were tested). (6) Protein solutions were introduced to sample wells (15 μL in each well). (7) Gels were run at 90 V through the stacking gel and then the voltage was increased to 120 V the resolving gel in the electrophoresis tank. (8) The gels were transferred to a small container containing a small amount of Coomassie blue (20 mL) and stained for two hours over a shaker. (9) Coomassie destain (about 50 mL) was added and agitated overnight to destain completely. (10) Photos were taken by using an imaging densitometer (Model GS-700, Biorad, CA, and USA) and gels were viewed for protein content.

Fat analysis

Fat content in whole milk was determined using the Babcock method as a non-solvent wet extraction method (AOAC 1995 a). In this method milk and concentrated sulfuric acid are mixed in a special bottle that has a long neck graduated to read fat percent. The acid digests the proteins and releases the fat. Warm water is added to the mixture to bring the fat into the neck of the Babcock bottle. The bottles were centrifuged, tempered at 60°C and then the amount of fat is read from the graduations on the bottle.

Fat content in fouling biofilm samples was determined using the Mojonnier method (AOAC 1995 b).This method uses ammonium hydroxide as a means of stripping the protein from the fat. A special Mojonnier extraction flask is used. It has a bulb on the bottom that is separated from the rest of the flask by a narrow neck. The aqueous phase settles to the bottom after centrifugation and water is added to bring the division between water and organic phases into the narrow neck. This allows one to pour off the ether-fat mixture quite accurately without pouring out any of the aqueous phase. The ether is evaporated leaving the fat behind. The fat is then weighed and percent fat calculated from the original sample weight.

Mineral analysis

The amount of mineral in the biofilms of different milk types was measured by analyzing the ash content of the biofilm (AOAC 1995 d). In this method a weighted amount of biofilm sample was placed in a crucible; the crucibles were placed in a muffle oven for 24 hours at 550 º C; and at the end they were reweighed to determine to the amount of ash. The percent ash was determined by dividing the amount of ash remained per initial dry weight of biofilm in the crucible.

RESULTS

<u>Effect of milk-type on biofouling (temperature method)</u>

Biofouling rates were measured over 6 ½ hours run time using two temperature measurement methods: (1) changes in Delta T (hot water inlet temp- milk inlet temp) and (2) decrease in milk outlet temperature. The fouling rate for each run was calculated from a linear regression of plots of temperature change versus time (Appendix A). Scatter plot was used to investigate the relationship between two changes in delta T and run time and also changes in milk outlet temperature and run time. Data is collected in x and y pairs by using the data logger to record one observation each 30 seconds for 780 observations and total run time of 6.5 hr. The best linear equation was determined for each milk type and also it was improved by taking median of 30-sample observation as demonstrated in Figure 5.1 and 5.2. Figure 5.1 show a positive linear relationship between delta T and time, and Figure 5.2 show a negative relationship between milk outlet temperature and time.
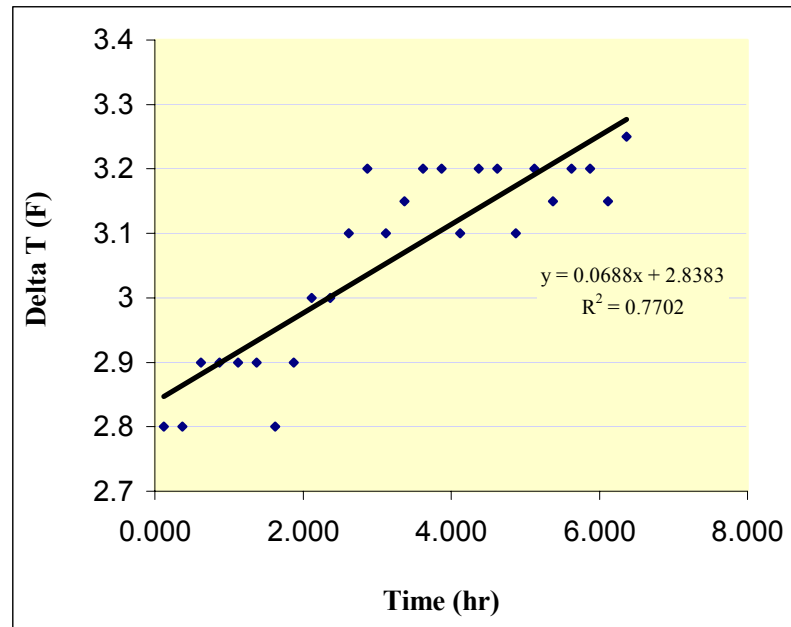
Figure R.1 Least square regression plots for control AB milk type
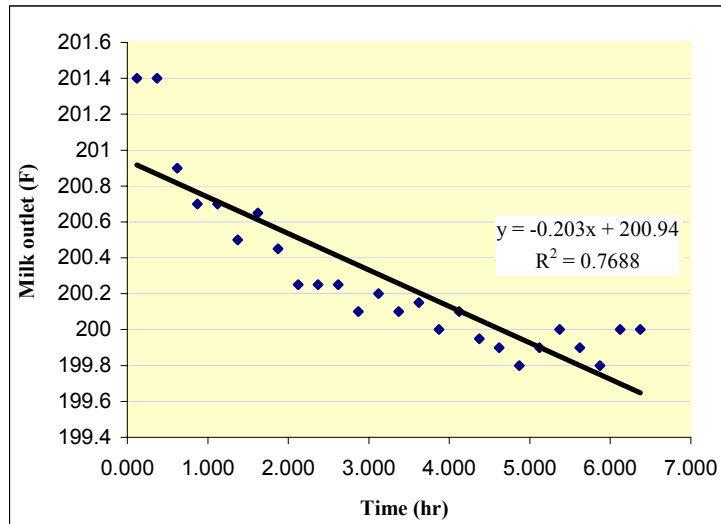
Figure R.2 Least square regression plots for control AB milk type

The three types of milk (Control AB, BLG BB and BLG AA) were analyzed with triple replications for each. Average and standard deviations of the measured biofouling rates for these three milk types are reported in Table 5.1 and 5.2.

Table R.1 Biofouling rate analysis based on a rise in delta T

Type	Test	Biofouling rate F/hr	Average Biofouling rate F/hr	Stdev
Control AB	1	0.0688	0.123	0.0513
	2	0.1294		
	3	0.1707		
BLGBB	1	0.02	0.0678	0.0583
	2	0.1327		
	3	0.0506		
BLGAA	1	0.1814	0.1382	0.0406
	2	0.1008		
	3	0.1324		

Table R.2 Biofouling rate analysis based on decreased milk outlet temperature

Type	Test	Biofouling rate F/hr	Average Biofouling rate F/hr	Stdev
Control AB	1	0.203	0.2877	0.0949
	2	0.2699		
	3	0.3903		
BLGBB	1	0.0631	0.133	0.0714
	2	0.2059		
	3	0.1299		
BLGAA	1	0.2375	0.2387	0.1017
	2	0.1377		
	3	0.341		

Results were similar for the two temperature methods: both methods showed a lower biofouling rate for BLG BB than for BLG AA and Control AB (Figure 5.3 and 5.4).
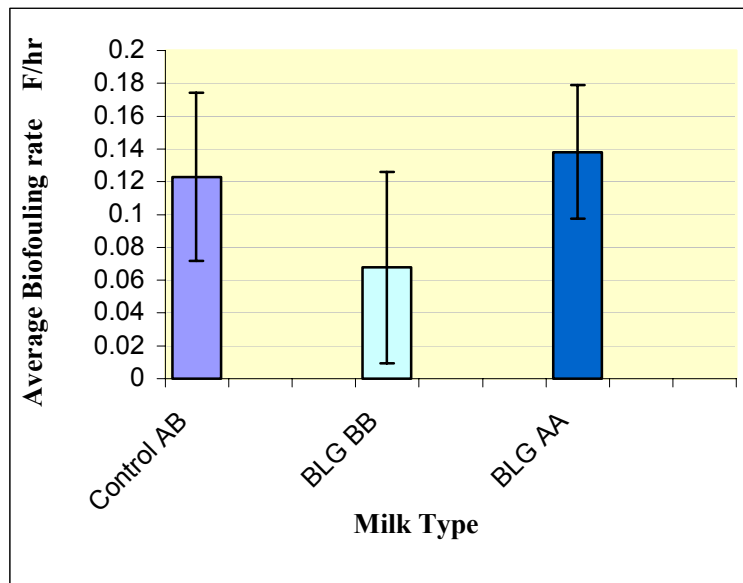
Figure R.3 Comparison of milk-type classification on fouling rate of the plate heat exchanger as a function of Delta T.
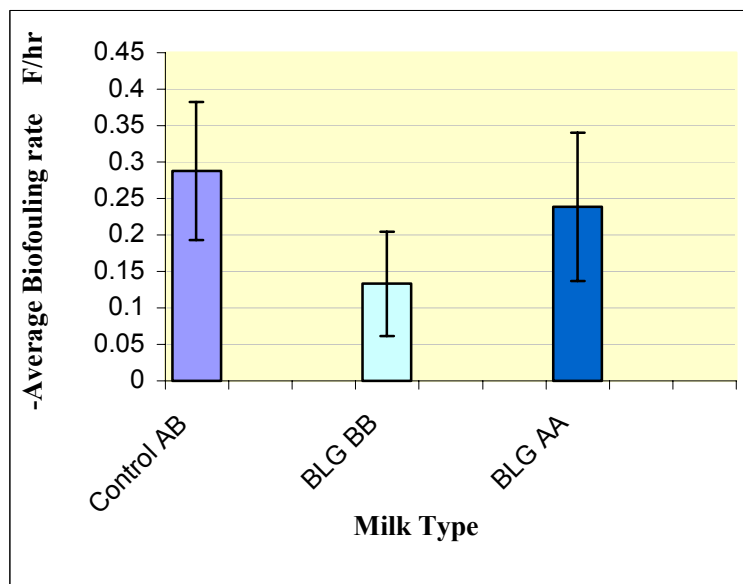
Figure R.4 Comparison of milk-type classification on fouling rate of the plate heat exchanger as a function of milk outlet temperature.

One-way analysis of variance by using Minitab and with considering 95 % confidences interval was employed to examine the differences among milk type (Table 5.3).

Table R.3. Statistical One-way ANOVA

Method	p-value
Delta T method	0.275
Milk outlet method	0.181

As indicated from Table 5.3 the milk outlet temperature method was resulted to indicate lower p-value, which suggests better response to biofouling rate for this method than Delta T measurement method. Overall, The AOVA shows the lack of significance of milk-type by using both methods at the 95 % confidence level.

Effect of milk type on biofouling (heat transfer method)

Overall heat transfer was calculated for different milk type over 6 ½ hours run time using energy balance equation between hot water and cold milk.

$$Q = (m \, C_p)_{milk} \, (T_{c2} - T_{c1})$$

$$C_p \, (milk) = 3706 \, \frac{J}{kg.\,^\circ K} \, \langle \, C_p \, (water) = 4186 \frac{J}{kg.\,^\circ K}$$

The heat transfer rate for each run was calculated from a linear regression of plots of Q (the amount of transferred heat) versus time (Appendix B). Scatter plot was used to investigate the relationship between these two changes. Data is collected in x and y pairs by using the data logger to record the milk inlet and outlet temperatures each 30 seconds for 780 observations and total run time of 6.5 hr. Milk mass flow was manually measured by using graduate cylinder and a lab timer. The best linear equation was determined for each milk type and also it was improved by taking median of 30-sample observation as given in Table 5.4.

Table R.4 Average transferred heat rate in the plate heat exchanger for different milk type

Type	Test	Q J/hr	Average Q J/hr	Stdev
Control AB	1	-95.292	-114.268	47.62892
	2	-79.052		
	3	-168.46		
BLGBB	1	-64.464	-159.288	115.17756
	2	-125.94		
	3	-287.46		
BLGAA	1	-84.298	-99.871	35.106394
	2	-75.245		
	3	-140.07		

Results of one-way ANOVA show lack of statistical significance difference between milk type in terms of a decrease in heat transfer rate, p-value was 0.621 at the 95% confidence level (Figure 5.5).
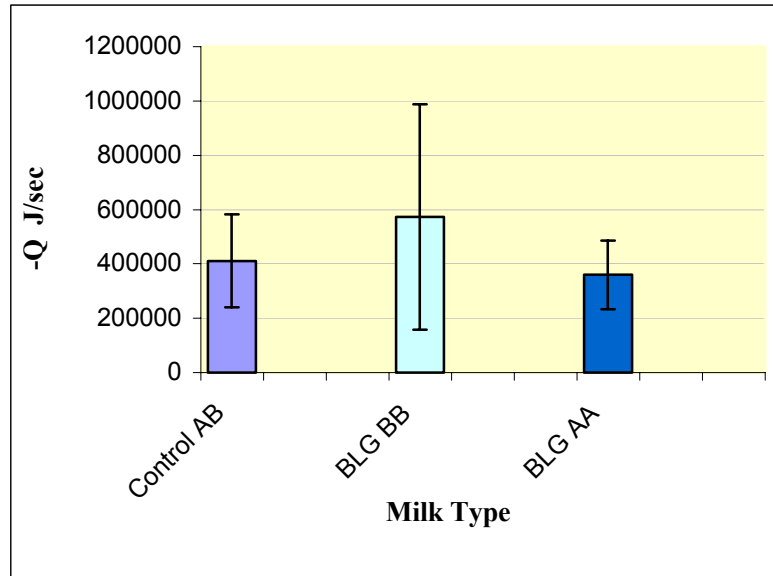
Figure R.5 Comparison of milk-type classification on heat transfer rate of the plate heat exchanger.

Effect of milk type on total amount of biofilm

Figure 5.6 shows the appearance of foulant material attached to the surface of the plate heat exchanger. The amount biofilm adhered to the surface of the heat exchanger plate was collected and analyzed (by scrapping off half the area from top to the middle of the first plate in the heating unit). Results are represented in Table 5.5 and Figure 5.7. About 20 % less total biofilm mass was observed for BLG AA than BLG BB or control BLG AB. However, the difference between milk types was not statistically significant at the 95 % confidence level and the p-value was 0.441 (analyzed by one-way ANOVA method).
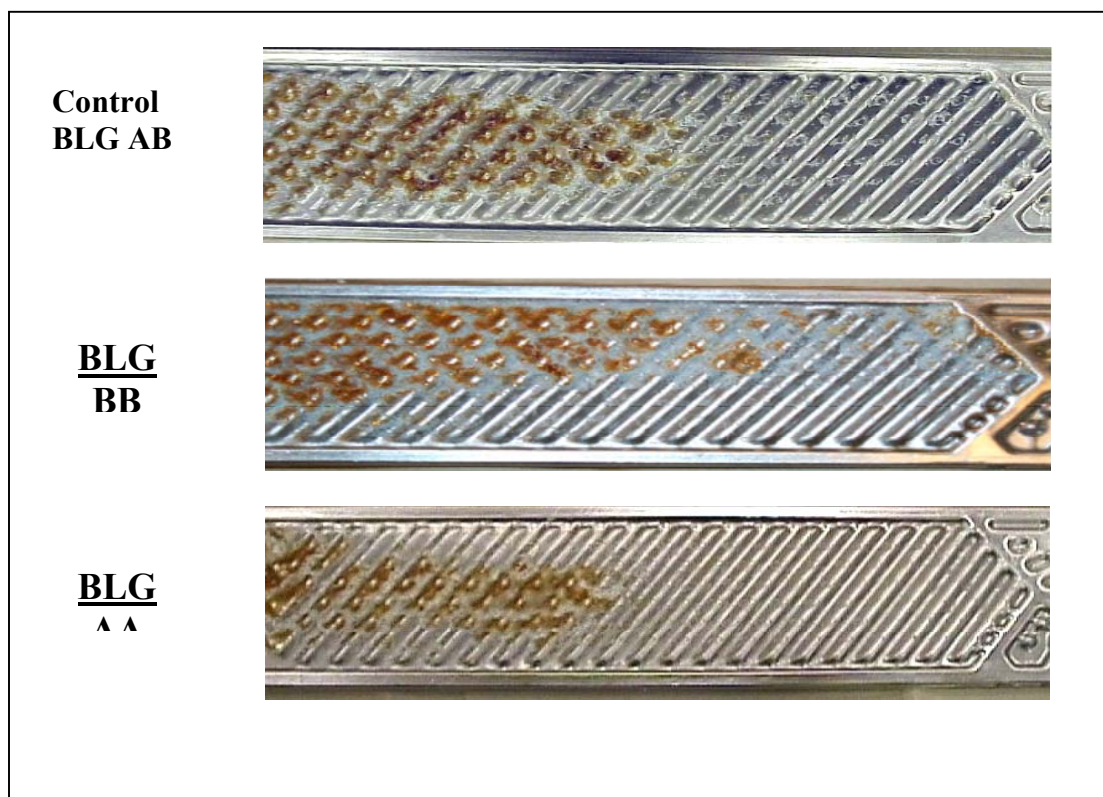
Figure R.6 Appearance of foulant materials deposited on plates; Control BLG AB, BLG BB and BLGAA

Table R.5 Dried biofilm mass analyses

Type	Test	Weight of dish g	Weight of dried biofilm+ dish g	Dried biofilm g	Dried biofilm g/m2	Average dried biofilm g/m2	Stdev
Control AB	1	12.2805	13.7875	1.507	100.47	93.18	6.32
	2	12.5928	13.9374	1.3446	89.64		
	3	12.5073	13.8486	1.3413	89.42		
BLG BB	1	10.5434	11.9664	1.423	94.87	90.59	29.33
	2	12.4555	13.3458	0.8903	59.35		
	3	12.3235	14.0868	1.7633	117.55		
BLG AA	1	12.4661	13.6038	1.1377	75.85	71.49	21.00
	2	12.5068	13.2366	0.7298	48.65		
	3	12.5418	13.8913	1.3495	89.97		
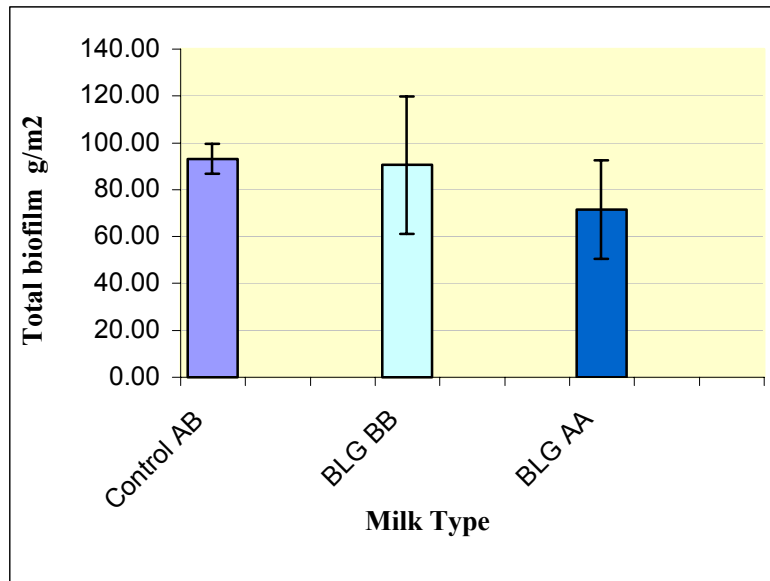
Figure 5.7 Quantity of total dried biofilm, collected on the surface.

Effect of milk type on protein content of biofilms

The average percent protein in dried biofilm samples was analyzed using the Kjeldhal method. Results are given in Table 5.6, and Figure 5.8 Greater percent protein was for BLG AA than BLG BB. The difference between BLG BB and BLG AA was statistically significant at the 95 % confidence with a p-value of 0.054 (analyzed by ANOVA method).

Table R.6 Average percent proteins in dried biofilms by Kjeldhal analyses

Type	Test	Weight of biofilm gr	HCl mL	% Nitrogen (w/w)	% Protein	Average Percent Protein	Stdev
Control AB	1	0.5055	21.9	6.07	38.70	37.49	0.89
		0.5086	21.9	6.03	38.46		
	2	0.5225	21.8	5.84	37.27		
		0.5135	21.4	5.83	37.22		
	3	0.5137	21.2	5.78	36.86		
		0.5121	20.9	5.71	36.45		
BLG BB	1	0.5158	21.4	5.81	37.06	35.79	0.93
		0.5037	20.8	5.78	36.88		
	2	0.4991	19.6	5.50	35.08		
		0.3699	14.7	5.56	35.50		
	3	0.5107	20.1	5.51	35.15		
		0.5065	19.9	5.50	35.09		
BLG AA	1	0.5013	22	6.14	39.20	38.48	1.05
		0.5142	22.7	6.18	39.43		
	2	0.5012	21.7	6.06	38.67		
		0.214	9.4	6.15	39.23		
	3	0.5174	21.4	5.79	36.94		
		0.5087	21.3	5.86	37.40		
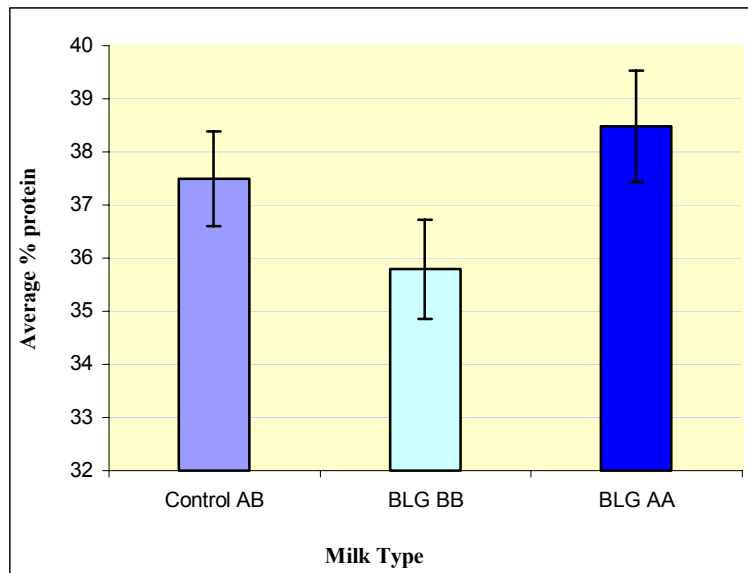
Figure R.8 Average percent protein (dry basis) in biofilms formed.

Figure R.9 represents the total protein content for three different milk types. Their trends are similar to the trend for total dry-weight of biofilm (Figure 5.7). One-way analysis of variance (ANOVA) reported a p-value of 0.509 for total amount of biofim. Therefore, no significant statistical difference was observed the 95% confidence interval.

By knowing quantity of dried biofilm samples (Table R.5) and their percent protein (Table R.6), the amount of protein can easily be calculated (Table R.7).
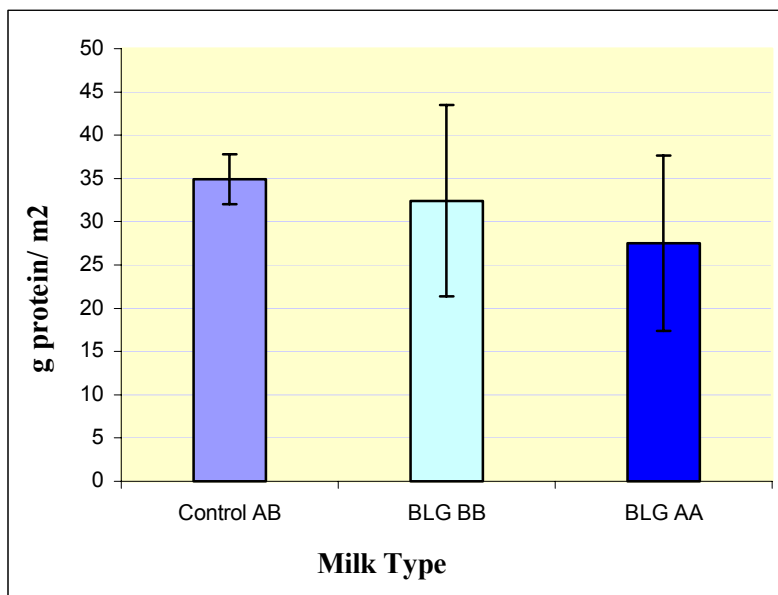
Figure R.9 Protein quantities per unit area of biofilm samples

Table R.7 Average quantity of protein in the biofilm

Type	Average dried biofilm g/m2	Average %Protein	Protein g/m2	Stdev
Control AB	93.18	37.33	34.78	2.90
BLG BB	90.59	35.61	32.26	11.06
BLG AA	71.49	38.27	27.36	10.15

Effect of milk type on the fat in biofilm

Fat content in fouling biofilm samples was measured using the Majonnier method. Results are given in Table 5.8. Greater fat content was observed for BLG AA type milk compared to the other type (BLG BB and the control BLG AB) (Figure 5.10). Analyses of these data with a one-way ANOVA resulted in a p-value of 0.095. The difference between BLG AA type milk and the two other types was statistically significant at the 95 % confidence level.

Table R.8 Analysis of fat content in biofilm

Type	Test	Weight of dish g	Weight of Sample g	Dried weight of dish + Fat g	Weight of fat g	% Fat	Average % Fat	Stdev
Control AB	1	27.9061	8.206	28.0039	0.0978	1.1918	1.1539	0.32
	2	41.732	8.1718	41.7987	0.0667	0.8162		
	3	30.2605	8.2005	30.3797	0.1192	1.4536		
BLG BB	1	30.3995	8.0092	30.5102	0.1107	1.3822	1.1705	0.26
	2	28.0679	8.4195	28.1417	0.0738	0.8765		
	3	29.9589	8.3018	30.0629	0.1040	1.2527		
BLG AA	1	33.964	7.795	34.0777	0.1137	1.4586	1.6793	0.23
	2	29.9544	7.8901	30.0853	0.1309	1.6590		
	3	28.0438	3.5254	28.1115	0.0677	1.9203		
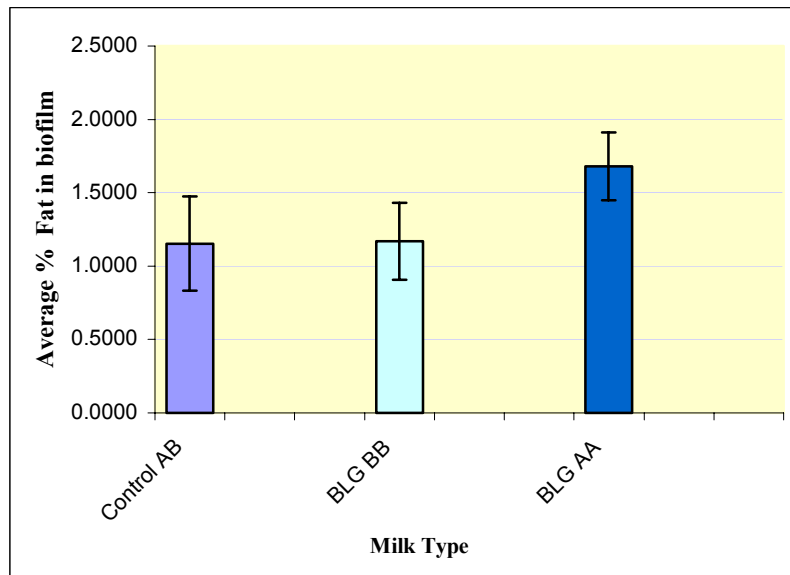
Figure R.10 Analysis of milk type on the fat in biofilm

Effect of milk type on the mineral in biofilm

Mineral content in biofilm samples was measured by using ash analysis (Table R.9). Figure R.11 shows the effect of milk type on mineral content in biofilm. Lower mineral content was observed for BLG AA than BLG BB or Control BLG AB. However, the difference between milk-type was not statistically significant at the 95 % confidence with a p-value of 0.447 (analyzed by ANOVA method).

Table R.9 Mineral content in biofilm

Type	Test	Crucible Wt. g	Wt. of samples g	Ashed wt. of cru + samples g	Wt. Of Ash g	% Ash	Average Ash %	Stdev
Control AB	1	21.0635	0.1126	21.0748	0.0113	10.0355	8.7955	1.29
		22.9429	0.107	22.9532	0.0103	9.6262		
	2	22.1221	0.1039	22.13	0.0079	7.6035		
		28.1103	0.1071	28.1195	0.0092	8.5901		
	3	18.8595	0.1074	18.867	0.0075	6.9832		
		23.831	0.1067	23.8416	0.0106	9.9344		
BLG BB	1	25.8714	0.104	25.8817	0.0103	9.9038	8.4510	1.80
		17.2819	0.1053	17.2897	0.0078	7.4074		
	2	28.2977	0.1052	28.3092	0.0115	10.9316		
		27.2234	0.1123	27.2336	0.0102	9.0828		
	3	18.7367	0.117	18.7441	0.0074	6.3248		
		22.9151	0.1063	22.9226	0.0075	7.0555		
BLG AA	1	25.9036	0.1297	25.912	0.0084	6.4765	7.3703	1.24
		24.564	0.1052	24.5698	0.0058	5.5133		
	2	27.3213	0.1124	27.3294	0.0081	7.2064		
		28.1572	0.114	28.166	0.0088	7.7193		
	3	17.3481	0.1055	17.3574	0.0093	8.8152		
		26.1574	0.1107	26.1668	0.0094	8.4914		
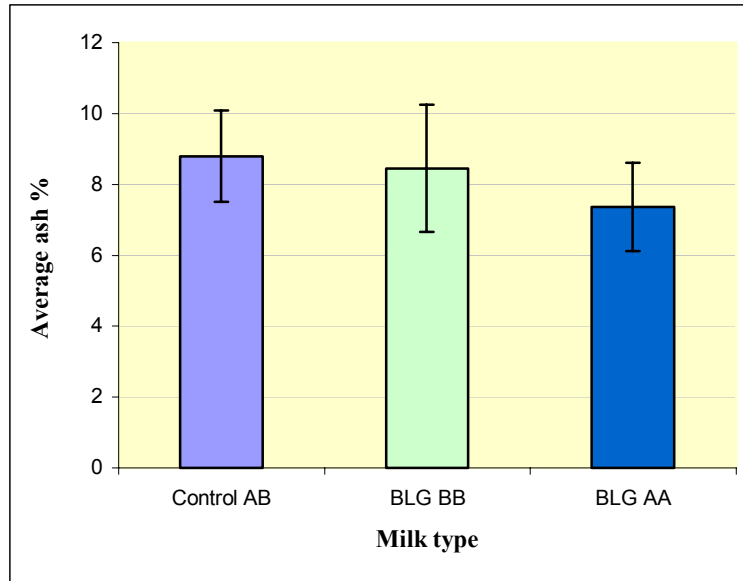
Figure R.11 Analysis of milk-type on the mineral in biofilm

LABORATORY SCALE RESULTS

The plates were consistently covered with biofilm as shown in Figure L.1

Figure L.Image of the first sample plate to come out from a ten hour run of type AB-AB raw milk. The plate was not completely covered with biofilm due to the milk outlet nipple was not at the very top. The exchanger was run in the position shown with the top of the exchanger up. This ensured that the water was in contact with the other side of the plate for the best heat transfer.

Elemental analysis of the plates used is represented in the next two tables:

Table I. These are the stainless steel sample compositions and their relative percentages. Manganese and sulfur are known to be minerals that can act as food for some bacteria.

	SS304	SS316	SS430
	0.08	0.08	0.12
Mn	2.00	2.00	1.00
Si	1.00	1.00	1.00
Cr	18.0-20.0	16.0-18.0	16.0-18.0
Ni	8.0-10.5	10.0-14.0	--
P	0.05	0.05	0.04
S	0.03	0.03	0.03
Mo	--	2.0-3.0	--
Fe	Bal.	Bal.	Bal.

Table II. This is the elemental make-up of the titanium metal sample. Note there should not be any sulfur of manganese in the metal system.

Titanium	
Al	5.50-6.75
V	3.5-4.5
Ti	Bal.
N	0.05 max.
C	0.1 max.
H	0.015 max.
Fe	0.4 max.
O	0.2 max.

Average of three trials
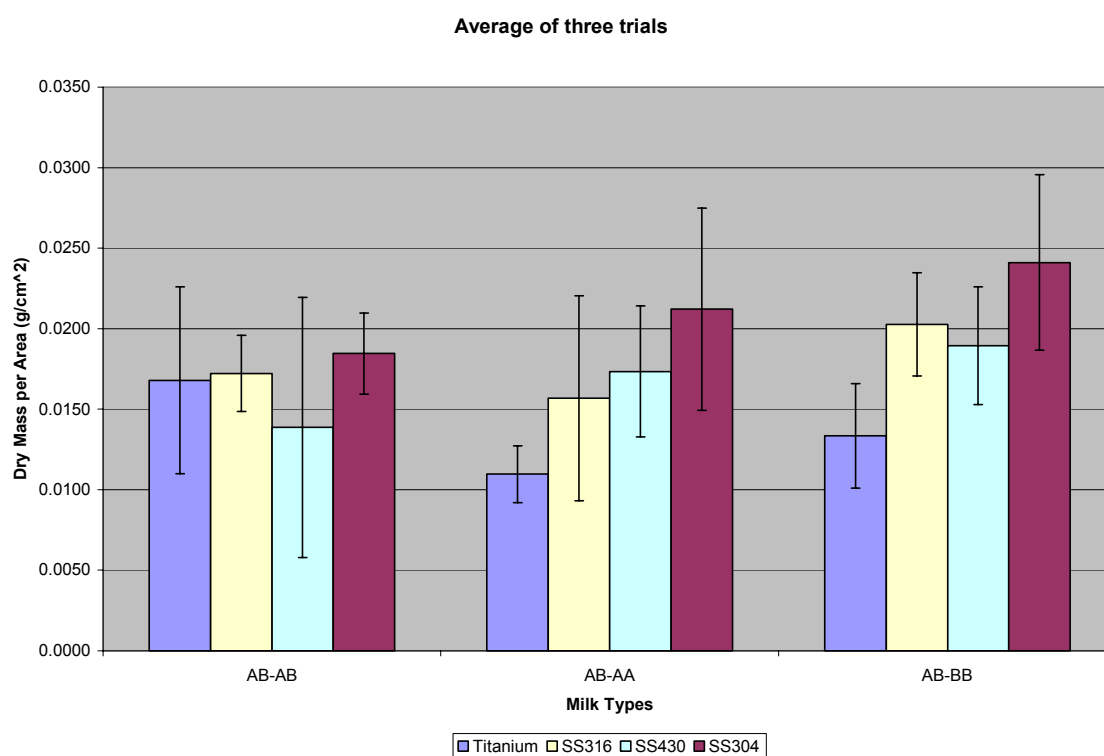
On average, titanium is the material that showed lowest degree of fouling, while stainless steel 304 showed a significantly greater tendency for fouling formation. This result allows us to focus on the composition differences of the materials. The most significant result regarding the milk is that the milk composition is the MAJOR contributor to fouling. Milks with high protein content do foul more, and the surface with significant more fouling was SS 304. Low protein milk in titanium showed least fouling.

Microscopy and Biofilm Analysis.

Using the SEM consistently yielded valuable images of the biofilm phenomena in

---

different scales.

Picture 1 demonstrates the ability to see at a 10X magnification the topography of the plates.
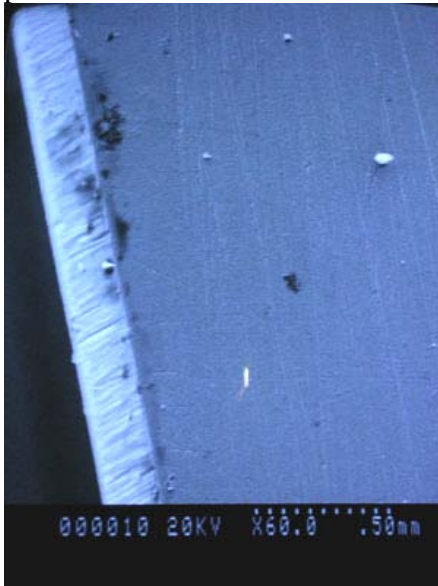
Picture 1. This is an SEM image of the edge of the titanium plate. The thickness is 0.040". The dark spots are that of a non conductive surface and represent either a thick oxide of biofilm. The white particles are probably dust particles.

It was also possible to obtain images of the macroscopic formation of aggregates of milk deposits. Pictures 2-4.
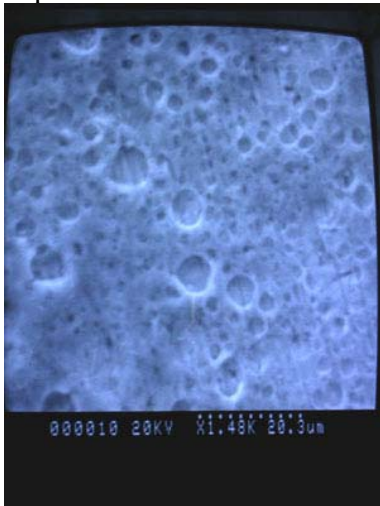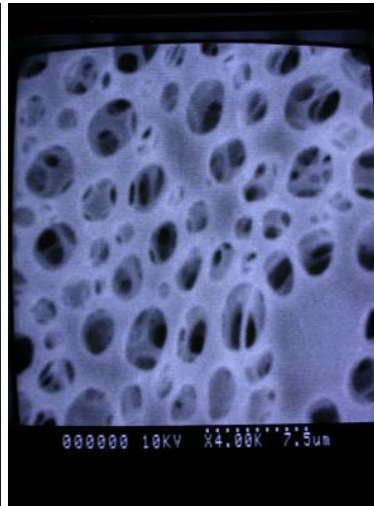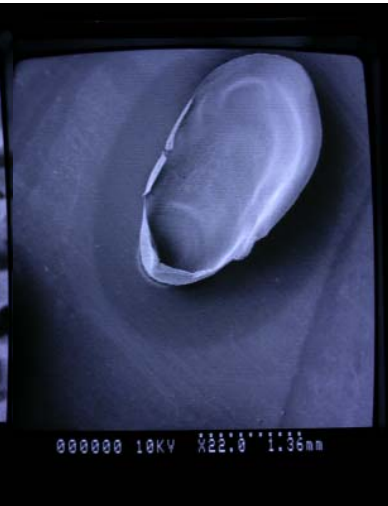
Picture 2   Picture 3   Picture 4

Most important for our purposes, it is possible to observe bacteria at the surface of the plates with a minimum of preparation. The results in pictures 5-7 demonstrate the principle.
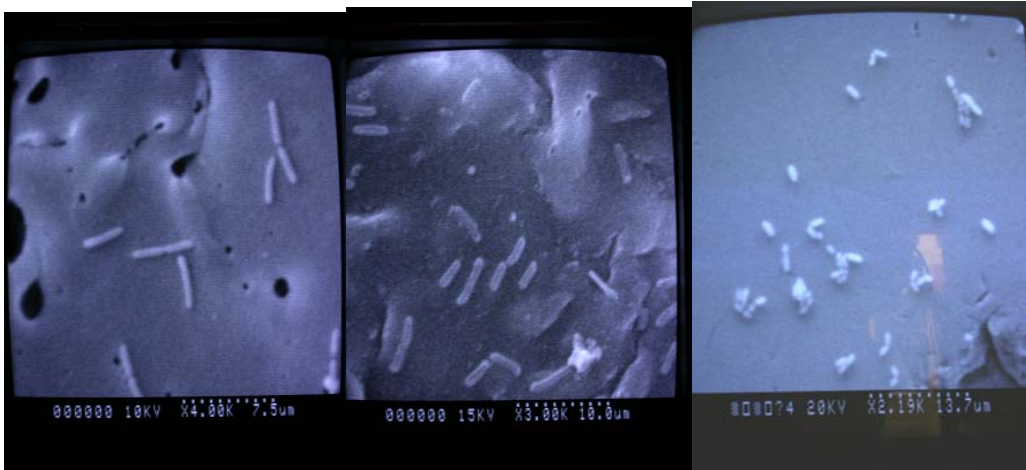
Picrue 5   Picture 6   Picture 7

CONCLUSIONS

The genetic variant milk with BLG BB produced the least biofouling in terms of loss of thermal conductivity in the heat exchanger and the total amount of biofilm produced by BLG AA milk on a dry-weight basis was lower than that produced by BLG BB milk, the percent protein and percent fat was lower for the biofilms produced by BLG AA milk. This suggests that the composition of the biofilm plays an important role in determining the severity of biofouling. Overall, this study confirmed that the BLG AA milk results in a small but statistically significant reduction in biofouling. The outcomes of this work tends to agree with the findings of New Zealand Dairy Board, in regards to lower fouling rate of BLG BB compare to BLG AA phenotype cows, with observing results by using linear regression delta T method.

The laboratory set up yielded data that allows us to conclude that there is a different and measurable interaction between the different types of milk and materials of the plate heat exchangers. Visual microscopic analysis proved that the main repository for bacteria can be found inside the structures formed by foulant (mostly protein and mineral). This suggests that initial fouling by milk components precedes bacterial deposition.

These findings accomplish most of the objectives originally established for this project, and indicate the high level of potential success of future data generated with these systems.

Development of a rapid method for detection of early spore germination using molecular probes

Rafael Jiménez-Flores
Dairy Products Technology Center

Christopher Kitts
Biological Sciences Department

Cal Poly State University

California Central Coast Research Partnership

Final Report, December 2003.

Title:
Development of a rapid method for detection of early spore germination using molecular probes.

Rafael Jiménez-Flores   Dairy Products Technology Center

Christopher Kitts   Biological Sciences, EBI

Summary

In this project we developed the basis for a rapid method to detect early germination of spores. The method is based on targeting the specific germination gene GerC3, which has been demonstrated to play a fundamental role in spore germination. In this work we proved that the probes designed exclusively detect spore-formers from the commonly found microbiological community found in milk powder. The method is similar to others that are used for sensitive detection of pathogens and virus. In addition we have developed a variety of traditional and novel techniques to detect and follow spore-former microorganisms and spores throughout commercial processing plants (using Terminal Restriction Fragment Patterns), which add to the wealth of information that support our work. We have used the basis of this method in commercial samples and found the validity of our results within the limits of normal microbiological environment of milk powder.

Major achievements:

- An endospore library of 5 distinct *Bacillus* was developed from our collection. These strains were selected for their multiple enzymatic activity and highly significant potential to reduce quality of milk.
- A comparison of different potential target genes was completed and tested for their suitability to the PCR based test.
- Design and test of 10 different sets of primers were tested for inclusion of all the selected strains in our library, and exclusion of commonly found bacteria in milk powder. A multivariate analysis resulted in optimized probes.
- Laboratory samples and commercial samples were tested using the probes resulting in satisfactory specificity.
- Sequencing analyses of the amplified fragments were performed to confirm the target gene, confirming target.

- A master's thesis was completed with data from this work, and 4 undergraduate students were involved in different specific objectives.

Note: The only limitation under current conditions is that the NCBI database for the genus Bacillus is only complete for *B. subtilis* and *B. licheniformis.*

Materials and Methods

Spore Making Procedure

From vegetative state, plated, induced to sporulate, harvested, washed, storage at 4°C indefinitely

1. Streak isolated cultures on TSA
2. Incubate at 35°C for 7-10 days
3. Scrape to remove cell growth
4. Place in 4°C overnight
5. Add 0.03% lysozymes, incubate 1 hour 37°C
6. Wash with 67 mM phosphate buffer plus 15 mM $MgCl_2$ (pH 6.8) to remove lysozymes by centrifuging and decanting the buffer, repeat 3 times.
7. Store at 4°C indefinitely   **Model strains**Five "BEA" Strains (broad enzymatic activity)

Presumptively identify the five DPTC strains that exhibit casein, starch and fat hydrolysis along with lactose fermentation

Primer design

The gerC primers were optimized following the sequencing of the gerC gene in *Bacillus licheniformis* (Rife 2002). SpoA primers were taken from the sequence of *Bacillus subtilis* found in GenBank (Accession Number NC_00964). The following primers were used. The pairs were optimized to find the most specific detection system.**Primer sets**–GRA/GFA (100)

```
For 5'-GGG TTT CNG TCT TCT TCA G-3'
Rev 5'-GGG GGC TAT TCG GCT CCG AGC-3'
```
–GRAB/GFA (500)
```
For 5'-GGG TTT CNG TCT TCT TCA G-3'
Rev 5'-GAA CCG AAA GCT CAT CGG-3'
```

DNA Extracion

Isolates from each DPTC culture were streaked on TSA for 24 hours at 35°C. *Bacillus* genomic DNA was extracted via Soil DNA isolation UltraClean™ kit (MoBio Laboratories, Inc., Solana Beach, CA).

Industry milk powder samples containing endospores were reconstituted to 10% solids and 3g of sodium citrate was added to chelate the calcium micelles, after a brief shaking period the samples were ultracentrifuged at 7000rpm for 30 minutes at 4°C. The supernatant was removed and replaced with 2.o mLs of DI water. 5.0 mLs of lysozyme solution (0.1 M potassium phosphate buffer, 15mM magnesium chloride, 0.9 ng/mL lysozyme) and 100mL proteinase K (Sigma) (0.020g in 1L water) were added to each sample and incubated at 35°C for 1 hour. The sample was centrifuged at 4000 rpm at room temperature for 20 minutes. The pellet was then used in the MoBio extraction kit.

PCR conditions Thermo-cycling conditions were as follows: an initial denaturation step of 94ºC for one minute followed by 35 cycles of 94ºC for forty-five seconds, 59ºC for forty-five seconds (54ºC for forty-five seconds for GRAB), 72ºC for one minute and fifteen seconds, and a final extension step of 72ºC for two minutes

Sequencing

B. Sequencing PCR product:

Using *Bacillus licheniformis* as a template, the gerC3 gene was amplified using the conditions specified in section above. The product was inserted into the pCR 2.1-TOPO® vector using the TOPO® TA Cloning Reaction (Invitrogen, Carlsbad, Ca). Blue/white colony screening and EcoR1 enzymatic digestion were used for the identification of a positive transformation. Samples were then delivered to Environmental Biotechnology Institute (EBI, California Polytechnic State University, San Luis Obispo, CA) for sequencing.

Samples were also sent to the DNA facility at Utah State University, which sequences DNA for researchers. The service is also available to off-campus researchers. The samples were ethanol precipitated before being sent. The precipitation consists of adding 50 μL of 100% ethanol plus 1μL sodium acetate to each sample. The sample is then placed in 4°C for 30 minutes followed by a centrifugation at 10,000 rpm for 10 minutes. The pellet is then washed with 50μL of 70% ethanol and spun as before. The pellet is then dried at 90°C for one minute. The samples are rehydrated with sterile water and frozen for shipment. Sequencing is performed an ABI prism 3100 Genetic Analyzer with Taq FS Terminator Chemistry (BigDye Terminators, dRhodamine Terminator and dGTP BigDye Terminators). Custom primers are to be provided by the user. The turnaround time for receiving sequencing results depends on the volume of sequencing requests to be processed. The electrophoretic data is returned as a computer-readable sequences file via e-mail. The sequencer is capable of yielding 800-900 bases of sequence per lane per run from a quality template/primer combination. The average sequence length is 450-600 bases with >98% accuracy (Ninglin Yin, Utah State University Sequencing Lab).

Analyzing the Sequence

The sequenced 100bp fragments were aligned with the original GerC sequence obtained from Utah State. The alignment was performed usuing the ExPASy SIM alignment tool (http://us.expasy.org/tools/sim-nucl.html). After confirmation of alignment in the correct region the fragments from each related families were aligned.

All sequences from related strains were compared and analyzed using dnaStar software Megalign program.

Results

•**Strain Classification**

Table 1 Strain identification **Species identification of the five "BEA"strains from the DPTC collection**

Bacillus licheniformis	ATCC# 14580
	SL3
	CL11
	CH3
Bacillus circulans	CL6
Bacillus subtilis	CL10
	ATCC#23857
Bacillus thuringiensis	ATCC#13366
	DPTC BOB1

••••PCR
500 bp primer set
The 300 bp Primer Set.

The *GerC3* sequence was obtained from the NCBI database for *B. subtilis, B. cereus and B. firmis*. In addition, a previously sequenced GerC3 gene amplified from a *B. licheniformis* strain isolated from a California milk powder was also used in primer development (Amy Rife, 2002). The sequences were aligned using Bioedit, a freeware program. The sequences were searched for conserved areas between the species. There is greater homology between *B. subtilis* and *B. licheniformis* in this gene than for the other two species. Therefore specific primers were developed specifically for these two species. The primers were chosen to be 17 base pairs long, with ~50% GC content, and Tm~ 50°C. They were designed to yield a 300bp product.

CowF 5'-3'
TGCGATGTATACAGGCG

CowR 5'-3'
TGATAAGACATGCCGAC

DNA was isolated from *B. subtilis 23857*, *B. licheniformis 14580*, and 5 strains of wildtype organisms isolated from milk powder, CH3, SL3, CL6, CL10 and CL11. DNA was also isolated from a wildtype strain of *B. thuringiensis*, referred to as Bob, isolated by the Biosciences department at Cal Poly State University, S.L.O. PCRs were also performed on DNA from control strains that do not form endospores, including *L. casei, En. faecalis, E. coli* and *Ps. flourescens*. All DNA extractions were performed using a MoBio microbial DNA isolation kit (Solana Beach, CA). A ready made PCR mix, Supermix Hi Fidelity (Invitrogen, Carlsbad, CA) was used in a volume of 45ul per reaction. For each reaction, 1ul of 10uM Cow F and CowR, 2ul of between 10-100ng/ul template DNA. The PCR was performed with an initial denaturation of 94°C for 5', and 35 cycles of 94°C for 1',

50.7°C for 1', 72°C for 2', and a final extension step of 72°C for 7'. The PCR was then held at 4°C, until placed at -20°C or until a gel was run.

PCR fragments were visualized in a 25ml 1.5% agarose gel in 0.5X TBE, ran at 85v for 45 minutes; 0.5ul of a 10ng/ml ethidium bromide was added to the gel for visualization of the PCR product. A 100bp ladder was used to determine the molecular weight of the fragment.

Once fragments were obtained, three of the isolates were chosen to be sent out for automated sequencing at Utah State University on an ABI prism 3100. PCRs were ran in triplicate as described previously for each strain, and then the fragments were cleaned using a MoBio PCR cleanup kit. Both forward and reverse strands were sequenced.

After the sequences were received, they were blasted against the NCBI database, to determine if the sequence was from the desired gene.

Results:
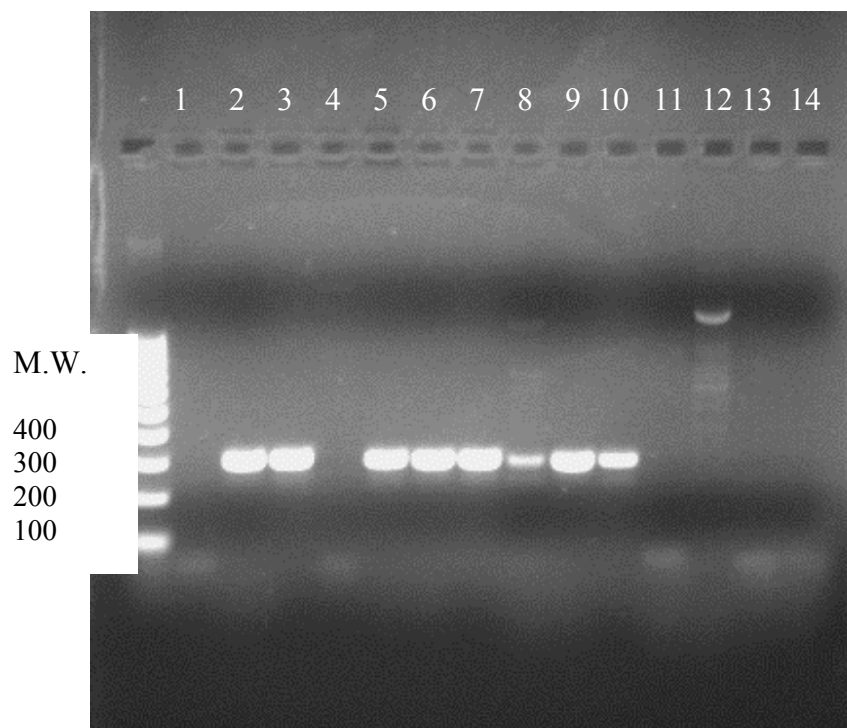
Lane #	ID	Lane #	ID	Lane #	ID
1	100bp ladder	6	SL3	11	L. casei
2	Open (-)	7	CL6	12	En. faecalis
3	B. subtilis 14580	8	CL10	13	E. coli
4	B. licheniformis 23857	9	CL11	14	Ps. flourescens
5	CH3	10	Bob	15	Closed (-)

PCR results show that 7 out of 8 of the *Bacillus* strains tested positive using the CowF/R primer set, yielding a 300bp fragment. Duplicate tests with CH3 concluded that this strain could not yield this fragment, suggesting a strain mutation within the selected primer sequence. The strain tested negative with all the non-endospore forming strands except *L. casei*. As a result, this test was repeated with three strains of *Lactobacillus*, including *L. casei 393*, *L. acidophilus 700396,* and *Lactobacillus sp. ATCC 53103* from human feces. These PCRs were run using freshly isolated DNA. Decontamination measures were taken on the equipment to ensure accurate results. The results of these PCRs yielded negative results suggesting that the positive result above was from *Bacillus* DNA contamination. Contaminated DNA is shown next to clean, fresh DNA for comparison. Lane number 13 and 14 show a failed reaction for strain CL11 due to low DNA concentration (diluted DNA). Undiluted DNA previously gave repeatable results for this strain.
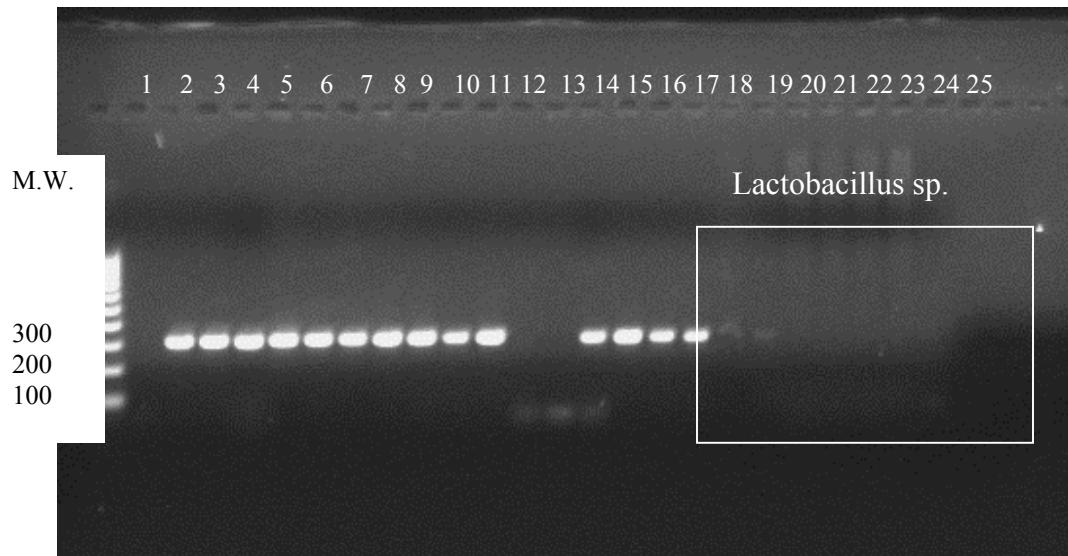
Lane #	ID	Lane #	ID	Lane #	ID
1	100 bp ladder	10	"	19	L. casei 393 (fresh)
2	Open (-)	11	CL10	20	"
3	B. licheniformis 14580	12	"	21	L. acidophilus 700396
4	"	13	CL11	22	"
5	B. subtilis 23857	14	"	23	L. spp. 53103
6	"	15	Bob	24	"
7	SL3	16	"	25	Closed (-)
8	"	17	L. casei contaminated		
9	CL6	18	"		

Sequences

Forward sequences contain the reverse primer sequence indicated in Bold (complement strand)
Reverse sequences contain the forward primer sequence indicated in Bold (complement strand)

B. licheniformis 14580

```
LL1-COWF-145_05.ab1.Seq  LENGTH: 344  Thu, Nov 20, 2003 9:56 AM  CHECK:
7181 ..
     1  TTTATGCTTG  CGGGATCTCT  TGAAATGATG  ACGAGAATTA  ACGAACCGAA
    51  AGCCCATAGG  ATTTTGTCAC  AGACGATCGT  TGAAGTTTGT  CTAGGGGAAA
   101  TTGAGCAGAT  CAAAGACAAA  TACAACATGG  AACAAAATCT  CAGAACGTAT
   151  CTCCGGCGTA  TCAAAAGAAA  AACGGCTCTC  TTGATCGCGG  TCAGCTGCCA
   201  GCTTGGTGCC  ATTGCGTCTG  GAGCTGATGA  GAAGATTCAT  AAGGCATTGT
   251  ACTGGTTTGG  GTATTAC**GTC**  **GGCATGCCTT**  **TATCA**
```

B. subtilis 23857

```
LL2-COWF-238_07.ab1.Seq  LENGTH: 339  Thu, Nov 20, 2003 9:56 AM  CHECK:
2728 ..
     1  TTTATGCTTG  CGGGATCTCT  TGAAATGATG  ACGAGAATTA  ACGAACCGAA
    51  AGCCCATAGG  ATTTTGTCAC  AGACGATCGT  TGAAGTTTGT  CTAGGGGAAA
   101  TTGAGCAGAT  CAAAGACAAA  TACAACATGG  AACAAAATCT  CAGAACGTAT
   151  CTCCGGCGTA  TCAAAAGAAA  AACGGCTCTC  TTGATCGCGG  TCAGCTGCCA
   201  GCTTGGTGCC  ATTGCGTCTG  GAGCTGATGA  GAAGATTCAT  AAGGCATTGT
   251  ACTGGTTTGG  GTATTAC**GTC**  **GGCATGTCTT**  **TATCA**
```

B. thuringiensis (Bob)

```
LL3-COWF-BOB_09.ab1.Seq  LENGTH: 341  Thu, Nov 20, 2003 9:56 AM  CHECK:
4188 ..
     1  TTTATGCTTG  CGGGATCTCT  TGAAATGATG  ACGAGAATTA  ACGAACCGAA
    51  AGCCCATAGG  ATTTTGTCAC  AGACGATCGT  TGAAGTTTGT  CTAGGGGAAA
   101  TTGAGCAGAT  CAAAGACAAA  TACAACATGG  AACAAAATCT  CAGAACGTAT
   151  CTCCGGCGTA  TCAAAAGAAA  AACGGCTCTC  TTGATCGCGG  TCAGCTGCCA
   201  GCTTGGTGCC  ATTGCGTCTG  GAGCTGATGA  GAAGATTCAT  AAGGCATTGT
   251  ACTGGTTTGG  GTATTAC**GTC**  **GGCATGTCTT**  **TATCA**
```

B. licheniformis 14580

```
LL4-COWR-145_11.ab1.Seq  LENGTH: 339  Thu, Nov 20, 2003 9:56 AM  CHECK:
6535 ..
     1  GTATACCCAA  ACCGTACAAT  GCCTTATGAA  TCTTCTCATC  AGCTCCAGAC
    51  GCAATGGCAC  CAAGCTGGCA  GCTGACCGCG  ATCAAGAGAG  CCGTTTTTCT
   101  TTTGATACGC  CGGAGATACG  TTCTGAGATT  TTGTTCCATG  TTGTATTTGT
   151  CTTTGATCTG  CTCAATTTCC  CCTAGACAAA  CTTCAACGAT  CGTCTGTGAC
   201  AAAATCCTAT  GGGCTTTCGG  TTCGTTAATT  CTCGTCATCA  TTTCAAGAGA
   251  TCCCGCAAGC  ATATAAT**CGC**  **CTGTATCCAT**  **CGCA**
```

B. subtilis 23857

```
LL5-COWR-238_13.ab1.Seq  LENGTH: 339  Thu, Nov 20, 2003 9:56 AM  CHECK:
6437 ..
     1  GTATACCCAA  ACCGTACAAT  GCCTTATGAA  TCTTCTCATC  AGCTCCAGAC
```

```
 51   GCAATGGCAC   CAAGCTGGCA   GCTGACCGCG   ATCAAGAGAG   CCGTTTTTCT
101   TTTGATACGC   CGGAGATACG   TTCTGAGATT   TTGTTCCATG   TTGTATTTGT
151   CTTTGATCTG   CTCAATTTCC   CCTAGACAAA   CTTCAACGAT   CGTCTGTGAC
201   AAAATCCTAT   GGGCTTTCGG   TTCGTTAATT   CTCGTCATCA   TTTCAAGAGA
251   TCCCGCAAGC   ATATAAT**CGC**   **CTGTATACAT**   **CGCA**
```

B. thuringiensis (Bob)

```
LL6-COWR-BOB_15.ab1.Seq   LENGTH: 341   Thu, Nov 20, 2003 9:56 AM   CHECK:
5056  ..
  1   GTATACCCAA   ACCGTACAAT   GCCTTATGAA   TCTTCTCATC   AGCTCCAGAC
 51   GCAATGGCAC   CAAGCTGGCA   GCTGACCGCG   ATCAAGAGAG   CCGTTTTTCT
101   TTTGATACGC   CGGAGATACG   TTCTGAGATT   TTGTTCCATG   TTGTATTTGT
151   CTTTGATCTG   CTCAATTTCC   CCTAGACAAA   CTTCAACGAT   CGTCTGTGAC
201   AAAATCCTAT   GGGCTTTCGG   TTCGTTAATT   CTCGTCATCA   TTTCAAGAGA
251   TCCCGCAAGC   ATATAAT**CGC**   **CTGTATACAT**   **CGCA**
```

Discussion:

All sequences when blasted against the NCBI database resulted in matches with the *B. subtilis 168* GerC3 gene sequence, confirming the value of using this primer set to isolate this gene fragment. The primers can now be further examined for use in Real Time PCR techniques.

It is important to note that there is inherent variability in the *Bacillus* genus, and that not all strains will produce this gene fragment using this primer set due to mutations that may occur in the primer locus. In the case of the reverse primer it is important to note that the sequences above are not an exact homology match when compared to the reference strains used to design the primers. This information could be useful in developing several reverse primers that could be used to more specifically target this gene sequence in a variety of *Bacillus* species. Degenerative primers may be more useful in compensating for the strain variability; however, degenerative primers may not be used in real time PCR techniques.

The remaining four *Bacillus* spp. that were not sent out for sequencing will be sent for automated sequencing at the EBI facility on the Cal Poly State University campus. The contaminated *L. casei* fragment will also be sent for sequencing to verify this contaminant. The sensitivity of this assay to detect *Bacillus* contamination on equipment may be useful in quality control audit systems, and validation of PCR research facilities.

It is recommended that an ATCC reference strain of *B. thuringiensis* be obtained for use in this study, as a reference strain may test differently than a wildtype strain of this species.

A blast alignment is included with this report as an example.

Conclusions

Potential Sequence Sites

Of genes focused on GerC showed positive results

  Specificity was satisfactory, and all model strains could be detected

  Useful for further analysis due to transcription early in germination

Primer set development

100bp most specific and easiest to interpret in PCR experiments

  Primers designed for RNA (RT-PCR) from these experiments demonstrated that can be used in a commercial operation.

Sequencing

  The sequences demonstrated homology to the GerC3 gene. Even though only two complete genomes for the genus Bacilli are currently available, the homology of the sequences demonstrates the principle. This confirms our hypothesis, and warranties great potential for the detection method. **Early detection**

300 bp, GerC3 primers enable specific detection of endospores in milk powder

  Identify spores before they can cause damage to product.

Increase possibility of "real time" detection in a variety of commercial samples. Specifically milk, but due to its complexity in composition it is easy to assume that it should work within other matrices.

In food systems

  Milk powder (food-borne illness)

In biological situations

  B. anthracis (human health) # References

1.   1. Levinson, H. *Sporulation and Germination*. in *8th international spore conference American society for microbiology*. 1980. Woods Hole, Mass.

2.   2. Caric, M., *Concentrated and Dried dairy Products*. 10 ed, ed. Y.H. Hui. 1994, New York, New York: VCH Publishes, Inc.

3.   3. Johnstone, K., *Pulling the trigger: the mechanism of bacterial spore germination.* Molecular microbiology, 1990. **4**(1): p. 137-141.

4.   4. Moir, A., *The Genetics of Bacterial Spore Germination.* Annual Review of Microbiology, 1990. **44**: p. 531-53.

5.   5. Byland, G., *Dairy Processing Handbook*, in *Dairy Processing Handbook*, T. AB, Editor. 1995, Tetra Pak Processing Systems AB: Lund.

6. Clement B., L.K., C. Kitts, *Terminal Restriction Patterns (TRFP's), a rapid, PCR-based method for the comparison of complex bacterial communities.* Journal of Microbiological Methods, 1998. **31**: p. 135-142

Access to the Central Coast's Marine Environment Through A Real-time/Archived Data Interface

Mark A. Moline

Biological Sciences Department
Cal Poly State University

Access to the Central Coast's Marine Environment Through A Real-time/Archived Data Interface.

FINAL REPORT

Mark A. Moline
Biological Sciences Department

Technical Summary

The purpose of this work was twofold: to demonstrate a real-time ocean observation capability and to provide data to address some of the environmental challenges impacting the local coastal region, such as hydrocarbon release, toxic algal blooms and runoff. The development of this capability is also aligned with the national Integrated Ocean Observation System (IOOS) that is partially sponsored by ONR through the National Ocean Partnership Program (NOPP; see http://www.ocean.us/ for more information).

Underside of pier +7.2 m(23.6') MLLW

26.2 m of cable from underside of pier to controller

7.2 m to underside of pier

MLLW level

Breakout for Wetlabs ECO-FLS Fluorometer

-1.5m MLLW — 3822 Sensor Disk with:
3930 Oxygen Optode, 3211 CT Sensor

-3.0m MLLW — 3823 fixture w/ 3211 CT Sensor

-4.5m MLLW — 3823 fixture w/ 3211 CT Sensor

-6.5m MLLW — 3823 fixture w/ 3211 CT Sensor

-7.5m MLLW — 3822 Sensor Disk with:
3712 Turbidity/Temperature, 3211 CT Sensor

Bottom -9.75m (32') MLLW

Figure 1. Configuration of Instruments

In support of these project goals, we have established a T1 communication connection to the Cal Poly State University marine facility at Avila Beach. Despite the remote location of the facility, the T1 connection is designed and engineered to exist within the present firewall of the Cal Poly State University. A pinhole through Cal Poly's firewall recognizes the connection, which is maintained by SBC Communications Inc. The T1 line is connected to a new SQL server on the pier that is mirrored by a second identical SQL server on campus. This provides the real-time capability (effectively no delay in data availability) as well as a redundant backup for data products. The SQL server at the marine facility has been directly connected to an array of in water sensors (**Figure 1**) and parses the real-time data stream for distribution.

The instruments package includes 5 temperature/salinity sensors, one turbidity sensor, one chlorophyll fluorometer sensor, one oxygen sensor and a petroleum fluorometer. The instrument package was installed in October, 2003 by divers and is configured to be easily retrieved to allow easy access for cleaning and calibration (**Figure 2**). The parsed data products (seen also my the mirrored server on campus) are now being archived on the server and also distributed real-time to a third web server that allows

Figure 2. Instrument chain being installed on Cal Poly's marine facility.

the user to visualize the products. A web site at www.marine.calpoly.edu has been developed for the live distribution of these data products. Embedded in the web server and the SQL server is one script that parses the data from the instruments for real-time use (**Appendix I**). A second script, a Graphical User Interface or GUI, then reads the parsed data for user defined visualization and data retrieval over the web (Appendix II). This GUI essentially has two portals; one for public outreach and one for more research driven institutions. Both are now available on the web to serve the public and allow federal, state, local agencies (and the public of course) to customize the real-time and/or archived data formats and products for their particular need (see **Figures 3, 4 and 5**). This customized real-time data system is unique and provides a testbed for this technology as related to coastal ocean data products. The entire project has benefited from the partnerships with Tenera Inc. (a local consulting firm) and Aandera Scientific.
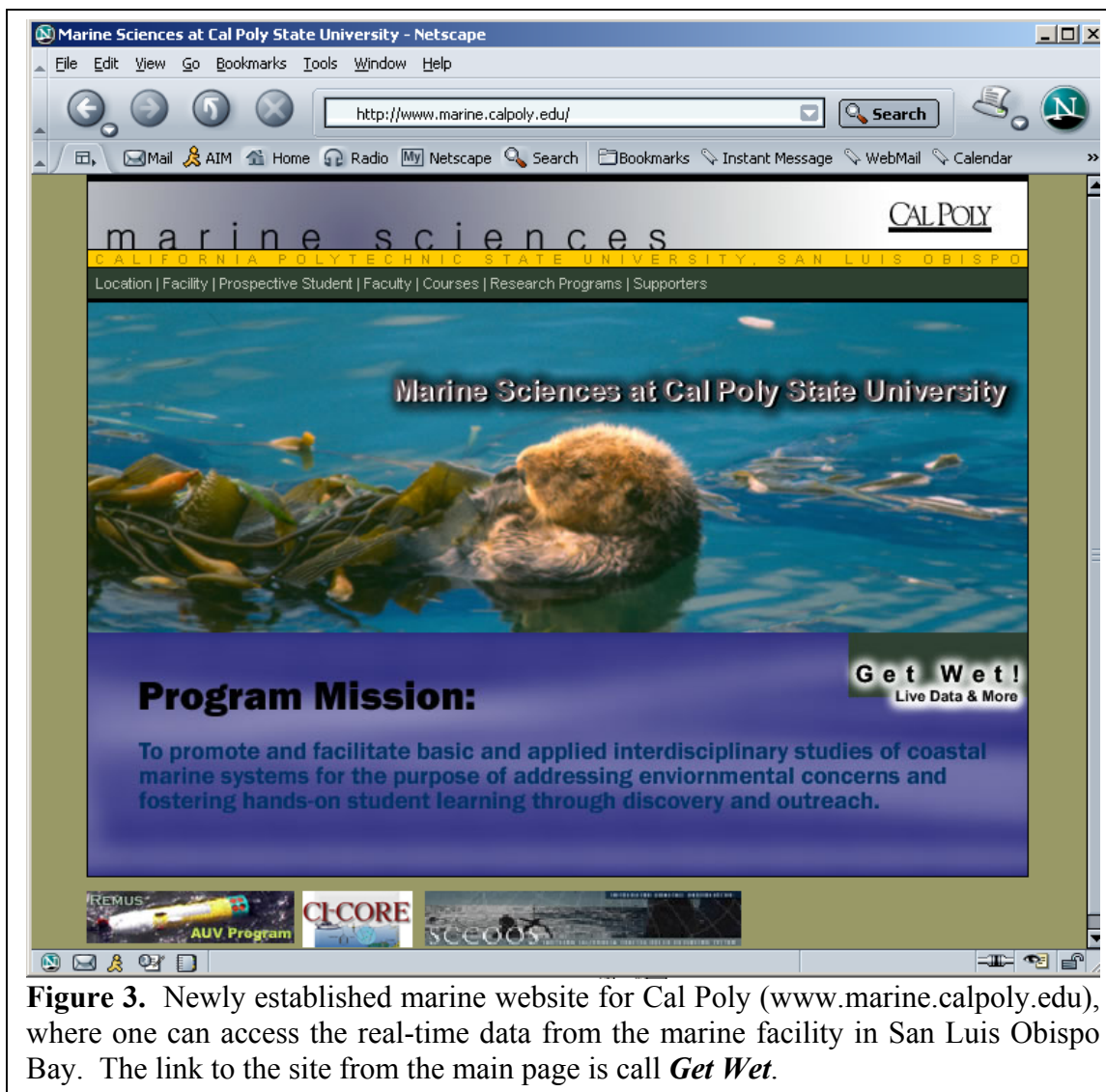
Figure 3. Newly established marine website for Cal Poly (www.marine.calpoly.edu), where one can access the real-time data from the marine facility in San Luis Obispo Bay. The link to the site from the main page is call *Get Wet*.
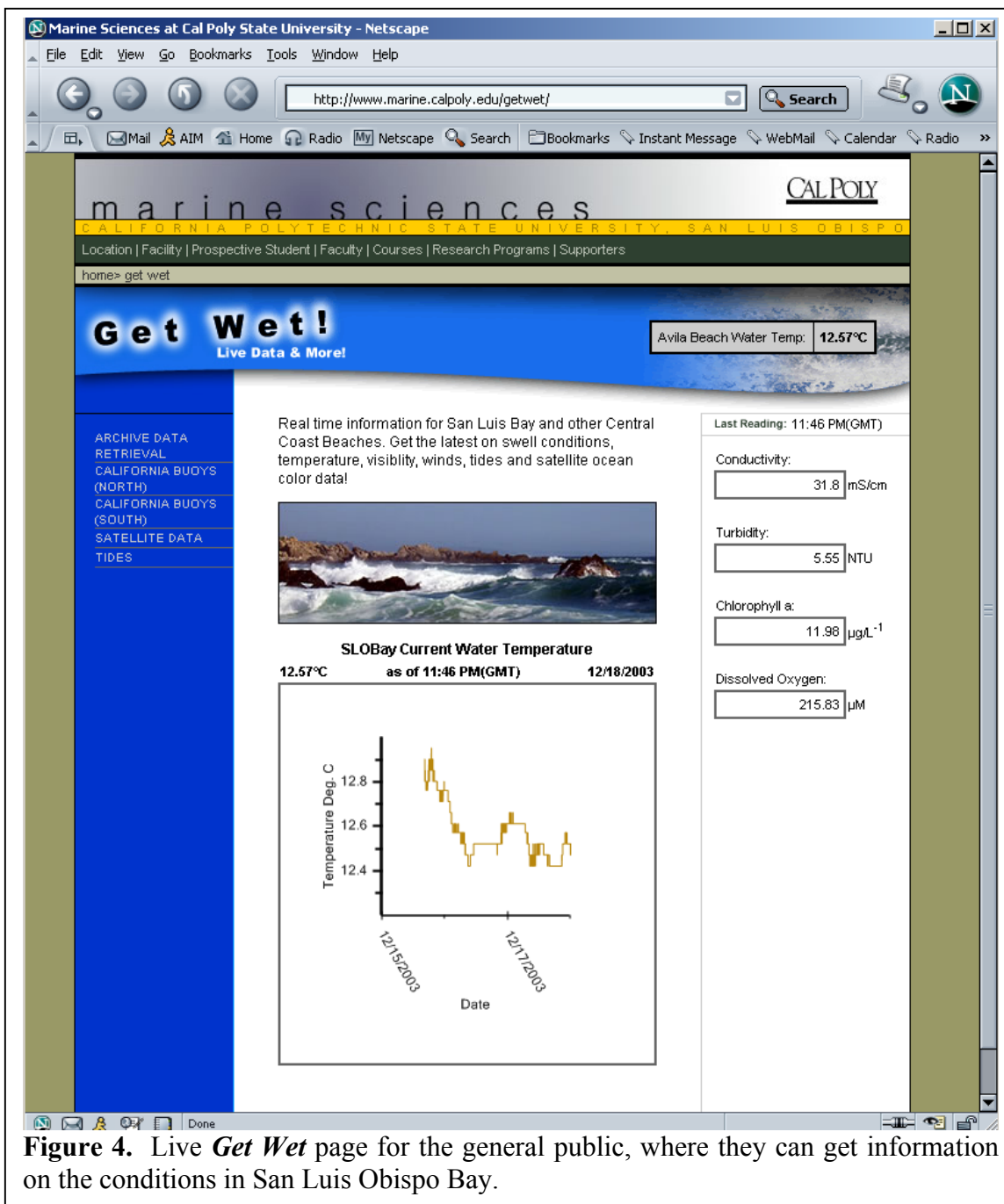
Figure 4. Live *Get Wet* page for the general public, where they can get information on the conditions in San Luis Obispo Bay.
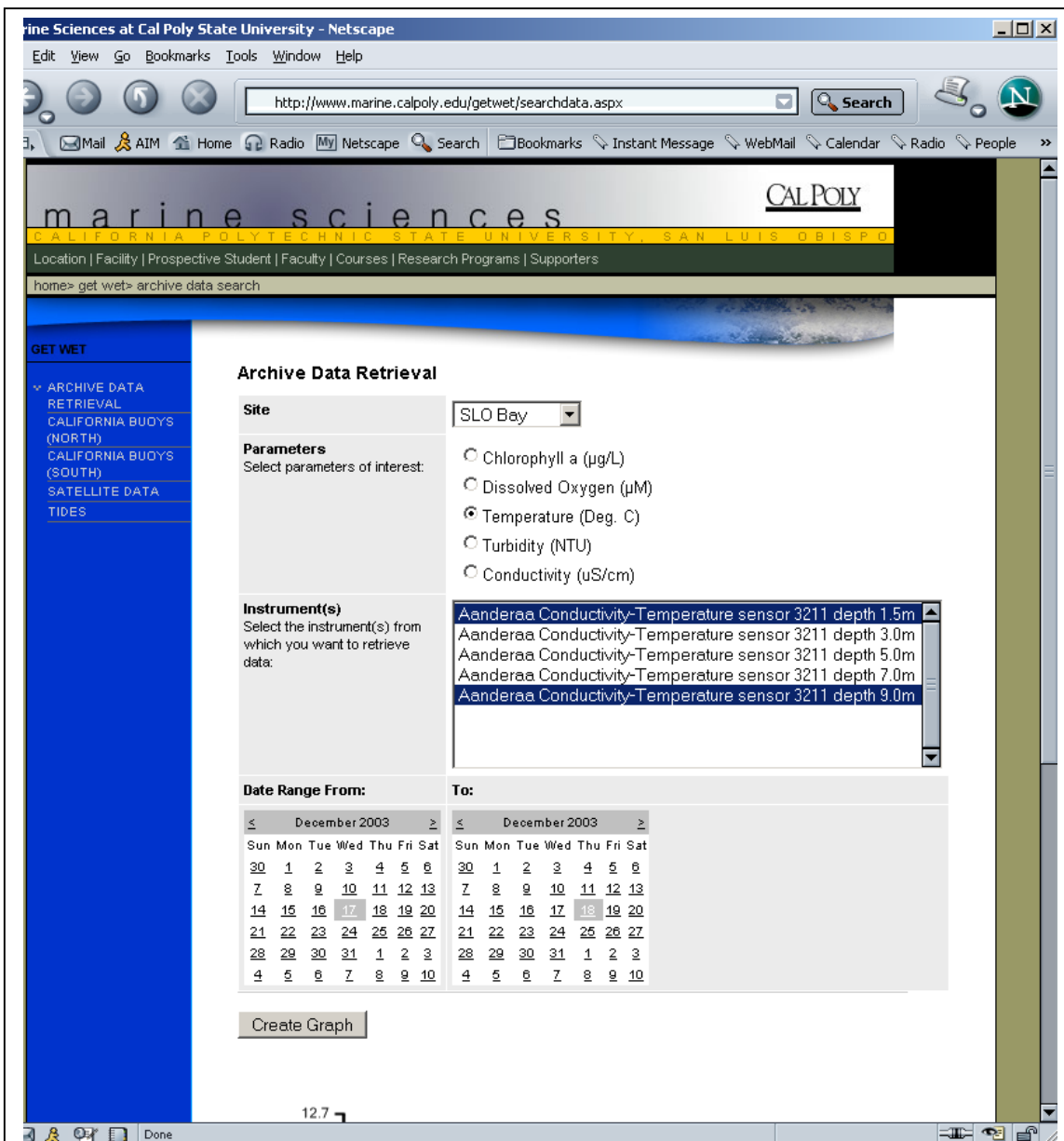
Figure 5. Live and archive data access page for research purposes and agencies, where they can get live and historical information on the conditions in San Luis Obispo Bay.

Impacts of Project

In Water Data

The real-time data stream is already being used by 3 students to examine local changes in physical hydrography and biological dynamics. Specifically this group of students is trying to quantify the degree of connectivity between the nearshore environment and the open ocean. Winds and currents are primarily generated from the northwest creating shadowing effects in northern embayments, such as the area near Avila where the marine station in located. We believe water gets isolated and trapped in these areas for a significant amount of time to develop a unique biological community that is significantly different than those a few kilometers offshore. To examine this and quantify the duration of this watermass isolation we are using the live data stream and comparing it to offshore buoys operated and maintained by NOAA. When the data sets converge, this would indicate a similar watermass and a flushing of the embayments. The rate of divergence of the two data sets will provide the time of isolation, which to interpret changes in the biological communities. Time-series measurements from this project are being supplemented by discrete water samples that are analyzed for phytoplankton composition to quantify success ional events in the context of the physical dynamics.

In addition to immediate student use, the local Regional Water Quality Control Board is using the data stream as a water quality indicator. Additionally, we are installing a petroleum sensor, which will give the Board an indication of hydrocarbon pollutants in the Bay.

Finally, the time-series station is leveraging the University in playing an active role in developing the national Integrated Ocean Observation System (IOOS; see http://www.ocean.us/). This system is rapidly developing and real-time measurements is one of the backbone technologies designed to provide more information about our impacted coastlines. From 2006-2012, the federal government is expected to fund this system in the neighborhood of about $600-800 M. Having existing infrastructure will give the University a voice in allocating these funds and implementing the system.

High Bandwidth Capability

The high bandwidth capacity developed by this project has numerous benefits and will help in further development of the marine facility and program.

- It will allow for multiple computer users at the pier. A 100-station router has been installed with plans to offer at least 24 stations within a wireless LAN at the pier. Students will have fast direct access from up to 2 miles away. This will not only allow flexibility while on the pier, but also would allow students to send and receive data while in boats around the Bay.

- There are plans to augment the instrument package already installed. This would include the installation of a weather station that would collect data at the same frequency as the in water measurements. This real-time data would be timely as the Navy's current high-resolution weather forecast model – the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) - is being tested and validated off the central coast of California. Currently, offshore NOAA buoys are used for validation, but several investigators at the Naval Research Laboratory at Stennis, MS and the Fleet's Meteorology and Oceanography Center (METOC) facilities in Monterey and San Diego have expressed interest in this project and having a coastal site. This data would also provide the context for the hydrographic data being measured.

- The high bandwidth also allows for data dense applications to be used. Examples of this include teleconferencing, web-cams for live local conditions as well as high-resolution underwater video, which could be used for a number of classroom activities. Remotely Operated Vehicles (ROV) have been used by the Navy for countless undersea applications. To date "remotely" has meant that a vehicle is hardwired to a control station that manipulates its movements and operations. The bandwidth at the marine facility now allows the possibility of extended operation of an ROV, from a cable connected to a control box to a web-based interface that potentially anyone could operate. The user interface would also allow students from Cal Poly State University and any other educational organization to use the ROV for learning and discovery.

Appendices Available Upon Request

238

Cal Poly Computer Network Performance Research Laboratory (NetPRL)

Phillip Nico
Hugh Smith

Department of Computer Science
Cal Poly State University

Cal Poly Computer Network Performance Research Laboratory (NetPRL)
C3RP Project #55380
Final Report

Phillip Nico   Hugh Smith

Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
email: pnico@acm.org
husmith@calpoly.edu

30 November 2003

1   Progress during 2002–03

We proposed to use the C3RP funds to continue the development of the Cal Poly Network Research and Performance Laboratory (CPNRPL), established last year by C3RP project 54410. This laboratory, now renamed the Cal Poly Network Performance Research Laboratory (NetPRL), is a research laboratory dedicated to investigating networking technology in conjunction with external partners from industry and academia.

As of Fall quarter, 2003, the laboratory comprises five participating faculty and ten graduate and undergraduate students from the Computer Science, Computer Engineering, and Electrical Engineering programs. Several other students have been involved in the project over the past year, but have either graduated or moved on to other projects.

During the past year we have continued to pursue our research into network performance and intelligent network interfaces while focusing on our long-term goal of developing NetPRL into a center of Operating Systems and Computer Networks expertise.

Some of this year's research accomplishments include:

- A paper, *An evaluation architecture for a network coprocessor,* written by J. Hatashita and J. Harris and H. Smith and P. Nico, was presented at the 2002 IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS) in November 2002. This paper describes work to develop an FPGA-based test platform for network coprocessors. A copy of this paper is attached.

1

- Another paper, *Toward a Common Host Interface for Network Processors,* was accepted for publication and presented at the Second International Conference on Communications, Internet, & Information Technology (CIIT) in November, 2003. This paper describes the current work undertaken this year to improve the interface between the CiNIC network processor and the operating system. A copy of this paper is attached.

- We now have an Internet 2 connection to the lab, allowing us access to its resources for experiments on throughput, latency, and quality of service (QoS).

- Two of our students conducted research into TCP/IP performance—including throughput, latency, and quality of service—by varying transmission line parameters. This work resulted in a senior project, *TCP/IP performance testing,* and led to a relationship with Ixia, a provider of network testing equipment.

- The Internet 2 connection, in conjunction with test equipment donated by Ixia has led to a partnership with Colorado State University to study network performance.

- This work, in turn, had led to developing a relationship with *Network Computing magazine.* We currently have one sneak preview product review article published, and two more in preparation.

In addition to these activities, we have continued our weekly seminar series in networking technologies, and have graduated a number of masters degree students.

At the same time as our research work, we have also continued to pursue relationships with industrial partners to gain support for the laboratory. The past year has been a difficult one for the technology industry, but nonetheless, we have enjoyed some successes:

- We traveled to San Jose to make a presentation on our work to to Brocade Communications Systems, which was met with interest.

- We traveled to Los Angeles to make a presentation on our work to Raytheon Company to make presentations on our work and to explore ideas for research partnerships. This resulted a grant of $1,000, and we currently have a proposal under consideration for $50,000.

- We submitted a proposal to Intel Corporation informing them of our work.

- We also made presentations to the Industrial Advisory Boards of the Computer Science, Computer Engineering, and Electrical Engineering programs. In all three cases the presentations were met with interest from each of the boards.

While we have not been successful in getting significant external financial support for the lab, we have reason to believe that our cultivation of these relationships will eventually bear fruit.

Next year we hope to continue developing the laboratory and seeking further industrial partnerships.

2

2 Future Work

During the Fall quarter of 2003, we have been mapping out plans for the future of the laboratory. In the coming year we hope to continue our work to develop the laboratory as a center of Operating Systems and Computer Networks expertise.

Our primary focus for the coming year will be around the development of a new test platform, a custom board, designed and produced by our students, with a CPU, FPGA, memory, and a PCI interface. This platform will allow us greater independence in our choice of implementation technologies and allow us to take better advantage of skills the students will learn in their coursework. The skills developed during the production of this prototype platform will transfer readily to other areas of network performance research.

3

Toward a Common Host Interface for Network Processors[*]

Eric Hawkins
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
ehawkins@calpoly.edu

Phillip L. Nico
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
pnico@acm.org

Hugh Smith
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
husmith@calpoly.edu

Abstract

Since their invention, network interfaces have generally been treated as special by operating systems because of their complexity and unique control requirements. In this paper we present a generic host interface through which an intelligent network interface or network processor can be managed as a simple networking device. To accomplish this, we push the complex network connection management and protocol processing code down onto the network interface. This new network processing platform is treated as a simple device by the host operating system. This model of host operating system interconnection provides for various network processor architectures to be handled identically using a well-defined kernel interface. Selection of the exact location for the kernel interface to the network processor was based on our goals to maximize the utility of the network processing platform, require no changes to existing network applications, and provide interoperability with existing network protocols (e.g. TCP, UDP). This paper documents the criteria and methodology used in developing such a kernel interface and discusses our prototype implementation using Linux kernel modules and our own ASIC-based intelligent network interface card.

Keywords: Information Systems and the Internet, Operating Systems Support, intelligent NIC, network processors

1 Introduction

The concept of offloading network processing from the host processor to a separate communication processor is not a new one. It has been discussed in the literature for some time, and several vendors have emerged to fill the newly created market niches for such devices. In order for such devices to be accepted into mainstream computing, however, a general interface is needed by which operating systems can offload network processing tasks to the co-processor without requiring a device-specific application programing interface (API) or other support mechanisms.

Traditional network interfaces have relied upon relatively dumb network adapters that simply connect the networking medium to the host I/O bus. These adapters are driven by networking code in the host operating system. To the host operating system, the network adapter looks like a small buffer into which packets can be written for transmission by the network interface card (NIC) onto the network. The Berkeley socket programming interface[8] is predominantly used by applications to send and receive data via the various network protocols supported by the operating system.

As networks have grown in size and complexity the network protocols have evolved to support these networks. Modern network protocol code is quite complex, especially protocols such as TCP/IP. Due to this growth in complexity, the processing required for transmitting and receiving data over the network has grown to a point where it is easy to justify the need for a device to offload these processing duties.[6][5] Much like the evolution of separate graphics processors was encouraged by increasing demands on graphics processing capabilities, the processing requirements of modern networking tasks are pushing the development of separate network processors for even the common network communication tasks.

Network processors provide many benefits beyond relieving host processors of common networking duties. Either through the use of general purpose CPUs or custom hardware, network processors can support auxiliary services right at the network edge. Security services such as intrusion detection or firewalling are enhanced by the physical separation of such devices from the host machine's software environment. Support for encryption can be incorporated into these devices either in hardware or software. Network quality of service mechanisms can be incorporated as well to enable multimedia applications.

Different approaches to the host-coprocessor interface have been proposed, but the most popular solution has been to use custom APIs and function libraries. The

---

Trapeze/Myrinet project[2] has shown impressive through-put across an intelligent network adapter, but it relies upon the Trapeze API to access the network adapter. Like-wise, the Nectar Communication Processor offloads pro-tocol processing as well as application specific tasks but does so through the use of the Nectarine programming in-terface which provides access to the Nectar message pass-ing facilities[1]. Network processors based on the Intel I2O specifications which utilize a split-driver model to iso-late host functionality from network interface functionality are also bound to a custom API for the host-coprocessor interface[3]. Since nearly all host-network processor inter-faces rely on custom APIs, the benefits of network proces-sors have not been realized on a broad scale. Incompatibil-ity with existing network software is a major impediment to the incorporation of these technologies.

To address the issue of binary compatibility we have defined an interface to the network processor that works along with the socket programming interface. We have de-veloped a prototype system that uses a well-defined Linux kernel interface at the top of the protocol stack. Using Linux kernel modules we have integrated support for the Cal Poly Intelligent Network Interface Card (CiNIC)[4] into the native Linux network protocol stack. Operations on CiNIC connections are dispatched through this interface in the host operating system to the CiNIC for protocol pro-cessing. Although the initial development has been done in Linux, the requirements and architecture of the interface can be applied to any operating system that supports the socket API.

The rest of this paper is organized as follows: In Sec-tion 2 we discuss the requirements of the host-network pro-cessor interface. In Section 3 we describe the kernel level interface selection for our prototype implementation. In Section 4 we describe our implementation and prototype platform. In Section 5 we discuss directions for future work. In Section 6 we present conclusions from our work.

2 Interface Requirements

We identified several requirements for the host-network processor interface. These requirements all stem from the primary requirement that our new OS interface be compat-ible with existing user-level interfaces so that existing pro-grams would not have to be altered.

Use socket API: Since the majority of legacy network ap-plication code uses the socket API, the interface to the net-work processor must exist within the scope of socket calls. Figure 1 shows the traditional socket call interface. The socket API sits at the top of the network subsystem and provides user programs with an interface to the operating system's network subsystem. The great majority of ex-isting network-based programs were written to the socket API, and so to ensure compatibility with existing programs the interface to the network processor must exist within the scope of the socket API. However, this requirement means
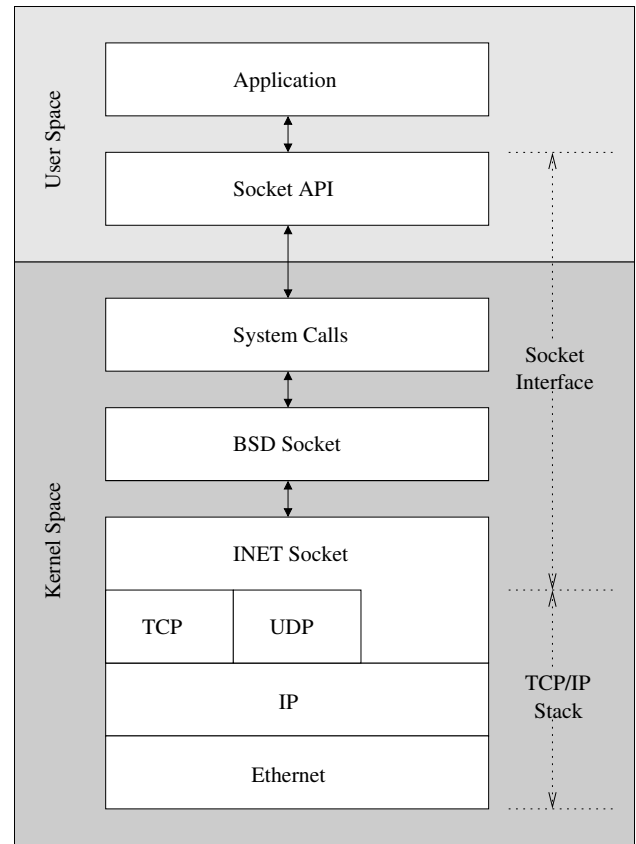
Figure 1. Socket API and OS interface.

that either the user-level socket library must be modified or support must be built into the OS. We chose the second ap-proach since OS modification is the best way to support the existing APIs and applications.

Utilize native device management: The network proces-sor should be managed as a normal device by the host oper-ating system. Custom application-level code should not be required to manage the device. Rather, existing operating system support for synchronization and protection should be used. The justification for this requirement is that pro-duction operating systems use proved access management policies and procedures. With a well-designed interface to the network processor, these existing mechanisms can be utilized intact. Requiring network processors to be treated as special devices outside the scope of operating systems' existing device management facilities poses a potential se-curity and robustness risk.

Look like a regular character device: To the host oper-ating system the network processor should appear to be a simple character device. The device should interact with the host operating system using a minimal number of oper-ations (open, read, write, close, etc.), and it should funda-mentally act like a simple device that raw data is written to and read from. This requirement is essential to preserving

the simplicity of the host-network processor interface. In implementation this requirement translates to locating the cleanest point at which to terminate host processing and invoke the co-processor. Determination of the point at which this division should be made is driven by two considerations.

First, the number of operations required for device support should be minimized in order to simplify the implementation of the interface. Second, the data structures used by the operating system for network devices must be kept consistent on both the host and the co-processor with minimum synchronization overhead. Many network related data structures are used at various points in the OS networking code and care must be taken to divide processing between the host and co-processor such that minimal replication of data is required. For example, the Linux kernel uses the *socket* data structure for filesystem information such as the inode number while the *sock* data structure is used for socket-specific network connection information. It is necessary to have the *sock* available to the co-processor since it takes care of network processing. This requires the *sock* to either be replicated between the host and co-processor or available only to the co-processor. On the other hand, the *socket* is required by the host who takes care of filesystem management. However, due to interdependencies between the data structures, separation of the two data structures would require synchronization overhead, but replication would require more.

3   The Socket Family Interface

As discussed in the previous section, the requirements for the host-network processor interface drove the design of the interface to be a kernel level modification. Several existing interfaces within the Linux kernel appeared as potential points to make the processing break between the host and co-processor.

The host-network processor interface could be implemented by intercepting all socket system calls destined for network processor connections and redirecting these calls to the co-processor. The host OS's system call table could be modified to redirect processing to functions capable of checking connections and dispatching calls appropriately. System call redirection minimizes the number of data structures requiring synchronization between the host OS and network processing platform. Also, due to the high level at which the processing division is made, system call redirection maximizes the amount of processing offloaded from the host to the co-processor. Unfortunately, the number of system calls and the requirement for catching all system calls makes this approach prohibitive in terms of implementation and execution overhead. The mapping mechanisms required to maintain network connections across multiple processes would also be complex and costly.

Another possible OS interface is the Virtual Filesystem Switch (VFS). The VFS is a software layer within the kernel that handles system calls related to a filesystem. It provides a common interface to various filesystems through specification of the functions and data structures that must be implemented to support a particular filesystem. The VFS seems like the natural spot to break host processing since it would allow network processor support to be implemented as a new filesystem type. Operations destined for the co-processor would be redirected through the VFS interface and handled on the co-processor. However, the implementation of OS socket call handling makes the VFS an inappropriate point for interfacing to the co-processor. The primary reason for this is that not all socket processing proceeds through the VFS. The *socketcall* multiplexer is actually a parallel data path to the VFS through which network operations can alternately be invoked. For example, to receive data from a socket, an application can make a *read* call, which is handled by the VFS implementation of the *read* system call. Alternately, an application can make a *recv* call on a connected socket, which is handled by the *socketcall* multiplexer and does not interact directly with the VFS.

The socket protocol family interface is a well-defined kernel interface just below the VFS layer. All socket processing converges from the VFS and *socketcall* multiplexer at the protocol family interface where it is dispatched to particular socket implementations. In the native networking code, this interface allows for the implementation of different socket types or protocol families. For example, with Internet domain sockets using the Internet Protocol (IP), this interface redirects socket processing to the set of data and operations defined for the Internet protocol family (PF_INET). The protocol family interface provides a narrow point at which to terminate host processing and invoke the co-processor on behalf of socket operations.

Using the socket protocol family interface, we have implemented a new type of socket family to be used with our network processor. We have named our protocol family PF_CINIC, since our prototype implementation utilized the CiNIC as previously mentioned. Figure 2 shows the software architecture of the network processor interface using the PF_CINIC protocol family alongside the host's native protocol families and network stack. Implementation of the PF_CINIC protocol family requires a minimal number of functions to be implemented (17 to be exact) due to the fact that the various possible data paths for socket operations converge at this point into the fundamental socket operations (e.g. *create, release, connect, bind, sendmsg, recvmsg,* etc).

Another advantage of making the break in host processing at the protocol family interface is that it provides a low-level view of the socket based only on the kernel data structures. Integration of the socket into the filesystem is handled by the kernel at a higher level, so all of the filesystem maintenance operations such as allocating and maintaining file descriptors for sockets are automatically taken care of. This allows the network processor to function as a true networking device without the overhead of filesystem operations, which would be required if host processing
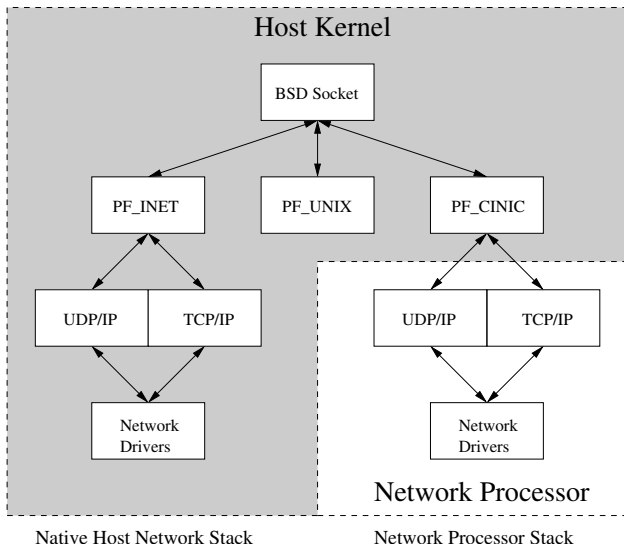
Figure 2. Protocol family interface

Figure 3. CiNIC development platform

was terminated at a higher level. The low-level view of the socket at the protocol family interface also limits the set of data structures affected by kernel operations on both the host and co-processor, providing for a clean separation of data between the host and co-processor with minimal synchronization requirements.

Breaking the host network processing at the protocol family level allows multiple network protocols to be supported by the co-processor. Protocol such as TCP/IP and UDP/IP are implemented at lower levels in the operating system, so processing destined for different types of sockets can proceed through the PF_CINIC interface and be switched based on the type of connection at lower levels. This approach stands in contrast to other so-called *offload engines* that can only support specific protocols.

4   The CiNIC Prototype Implementation

Our prototype implementation utilizes our Strong-ARM processor based CiNIC for the network co-processor. It runs a full Linux 2.4 operating system and is connected to the x86 Linux 2.4 host computer through a PCI-to-PCI bridge. A shared memory interface between the host and co-processor provides communication between the two platforms. Figure 3 shows the development platform.

Socket operations destined for CiNIC connections are intercepted at the protocol family layer, just at the top of the network protocol stack. The implementation of the PF_CINIC interface uses Linux loadable kernel modules, which are loaded prior to network processor usage much like a standard device driver. When host processing reaches the PF_CINIC interface, a communication packet is built with the necessary data and control information to transfer processing to the co-processor. This communication packet is then placed onto shared memory and a data ready
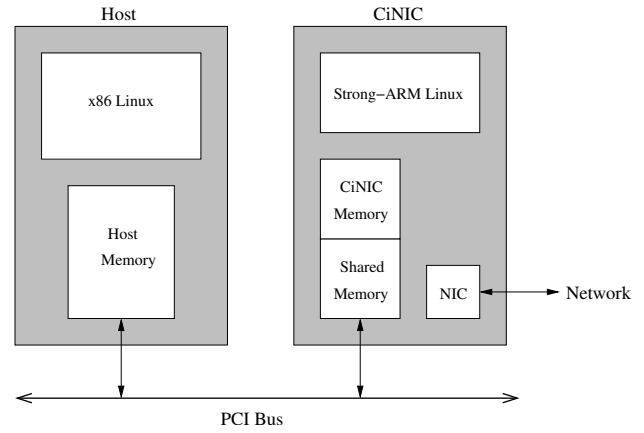
flag is marked. The host process is then put to sleep until the call returns from the co-processor. Figure 4 shows the shared memory communication protocol and threading architecure for the host-network processor interface.

A kernel thread on the co-processor is responsible for retrieving the communication packet from shared memory. The current implementation uses a polling protocol in which a kernel thread constantly checks the data ready flag in shared memory to see if a communication packet is ready for handling. When a communication packet is ready, this thread moves it out of shared memory and clears the flag. The communication packet is then placed on a wait queue to be handled by another thread responsible for maintaining socket connections and dispatching processing on each socket to worker threads. This *handler thread* allocates and deallocates socket data structures. It maintains a mapping of host sockets to co-processor sockets so that subsequent socket calls on a pre-existing socket proceed through the proper connection. The *handler thread* also manages a pool of *worker threads*, which are responsible for all socket processing other than creation and destruction. These *worker threads* pick up where processing left off on the host, calling functions within the native network family implementation.

When the co-processor completes its work on behalf of a socket call, all return data and control information are in the associated communication packet, which is placed onto shared memory for the host to pick up. The host utilizes a kernel thread that polls shared memory for returning packets. When a communication packet arrives, the host thread pulls the packet out of shared memory and wakes up the associated sleeping process. The native host process resumes where it left off and the socket call returns.

5   Future Work

Several optimizations are possible for the host-network processor interface as well as for our prototype platform. We are currently working to reduce the number of data

Figure 4. Shared memory communication architecture.

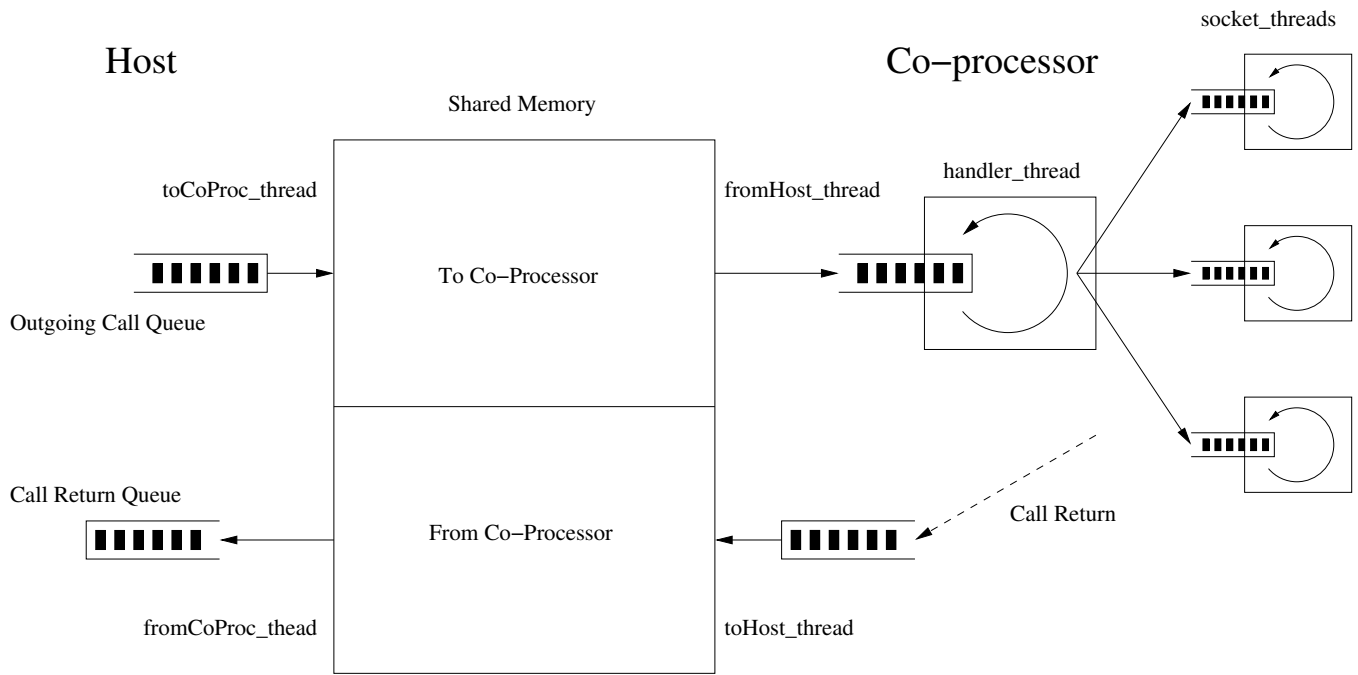copies required for communication between the host and co-processor in our development platform. The current implementation requires data to be copied from the host to shared memory and from shared memory to the co-processor (and vice versa). Since data copy operations are a major bottleneck in network processing[7], we need to reduce the number of copy operations in order to get reasonable performance. However, unlike other performance-oriented research on network processors, our goal is not to enhance overall network performance, rather to provide a standard interface to the network processor through which the network processor can provide various processing tasks.

Eventually, we plan to investigate how our Linux implementation of the host-network processor interface ties into the structure of networking code in other operating systems. We expect that other Unix implementations should coincide fairly well with the Linux implementation. The correlation to other proprietary operating systems may not be so close.

We are also developing a next-generation CiNIC using a FPGA design with an embedded soft-core processor running Linux. This future platform will provide us with many hardware capabilities beyond that of the current Strong-ARM platform such as the ability to create auxiliary processing blocks for special purposes. Along with development of our next-generation hardware, we plan to move from a polling communication protocol between the host and co-processor to an interrupt-driven communication protocol. This approach will be facilitated by the new hardware and will relieve both the host and co-processor

from supporting the busy-waiting kernel threads used in the polling protocol.

6 Conclusions

We have described the implementation of a host-network processor interface that relies upon the traditional socket programming API. We implemented the interface in kernel space using loadable Linux kernel modules. The selection of the network protocol family fulfilled our design requirements for the host-network processor interface by providing a narrow point at which to terminate host processing. This interface allows network processing to proceed on the outboard platform with minimal synchronization overhead, and allows the network processor to look like a simple device to the host operating system.

References

[1] COOPER, E. C., STEENKISTE, P. A., SANSOM, R. D., AND ZILL, B. D. Protocol Implementation on the Nectar Communication Processor. In *Proceedings of the SIGCOMM Symposium on Communications Architectures and Protocols* (1990), ACM, pp. 135–144.

[2] GALLATIN, A., CHASE, J., AND YOCUM, K. Trapeze/IP: TCP/IP at near-gigabit speeds. In *Proceedings of the USENIX '99 Technical Conference* (June 1999), pp. 109–120.

[3] GUTLEBER, J., AND ORSINI, L. Architectural software support for processing clusters. In *Proceedings*

of the IEEE Int'l Conference on Cluster Computing (2000), pp. 153–161.

[4] HATASHITA, J., HARRIS, J., SMITH, H., AND NICO, P. An evaluation architecture for a network coprocessor. In *Proceedings of the 2002 IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS)* (November 2002).

[5] KAISERSWERTH, M. The Parallel Protocol Engine. *IEEE/ACM Transactions on Networking 1*, 6 (1993), 650–663.

[6] KANAKIA, H., AND CHERITON, D. R. The VMP network adapter board (NAB): High-performance network communication for multiprocessors. In *Proceedings of Sigcomm-88* (1988), pp. 175–187.

[7] KLEINPASTE, K., STEENKISTE, P., AND ZILL, B. Software support for outboard buffering and check-summing. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication* (1995), pp. 87–98.

[8] STEVENS, R. W. *UNIX Network Programming*. Prentice Hall, Englewood Cliffs, NJ, 1990.

An Evaluation Architecture for a Network Coprocessor[*]

Jason Hatashita and James Harris
Department of Electrical Engineering
California Polytechnic State University
San Luis Obispo, CA 93407
email: jhatashi@alumni.calpoly.edu
jharris@calpoly.edu

Hugh Smith and Phillip L. Nico
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
email: husmith@calpoly.edu
pnico@acm.org

Abstract

This paper outlines research currently being conducted by the Cal Poly Intelligent Network Interface Card (CiNIC) project to develop a network coprocessor. The purpose of this coprocessor is to free the host machine from its network processing duties as well as to allow for additional functionality such as hardware-based firewalling or quality of service (QoS) support.

We provide an overview of the current CiNIC architecture as well as an introduction to and evaluation of the next generation CiNIC architecture. Our evaluation consisted of analyzing the performance and capabilities of an FPGA processor in order to determine whether it will meet our future development requirements. The FPGA's performance was tested by timing the execution of the uClinux TCP/IP stack during send operations. The processor's capabilities were tested by adding custom logic to the system and interfacing it with the uClinux operating system. We determined that both the performance and flexibility of the FPGA make it an ideal next generation CiNIC architecture.

Keywords: Network Interfaces, Network Performance, Reconfigurable Architecture, Network Coprocessor

1 Introduction

In recent years, computer systems have become more and more reliant on networks in the performance of their most common tasks. The increase in network communications has brought with it a corresponding increase in the computational burden on the processor to support them. At the same time, microprocessors have become less expensive and more powerful leading many to consider the potential benefits of adding a dedicated network coprocessor to free the host machine's resources for more important work. This goal applies equally to true parallel architectures[8] and networked workstations[7].

The goal of the Cal Poly Intelligent Network Interface Card (CiNIC) project is to offload the network functions from a host machine onto a dedicated network coprocessor. In this case, we define network functions to be those processes that manage the movement of data to and from the network, including TCP/IP stack processing, firewalls, routing, and Quality of Service (QoS) features. Our current development architecture consists of a x86 Linux host machine and a STRONG-ARM Linux coprocessor which sits in the host's PCI bus and appears to the host to be a simple PCI device. The system architecture is shown schematically in Figure 1.

In spite of its simple appearance, the CiNIC coprocessor is a complete linux system in itself. Running a full operating system on our coprocessing board provides developers with a familiar, easy-to-use, environment and affords access to industry-proven software for the coprocessor's network functions. The flexibility of a microprocessor and full operating system on the coprocessor system allows for experimentation with any of the host's software routines. Rapid software prototyping and easy access for performance evaluation make this architecture a powerful and flexible platform for exploring the capabilities of a network coprocessor.
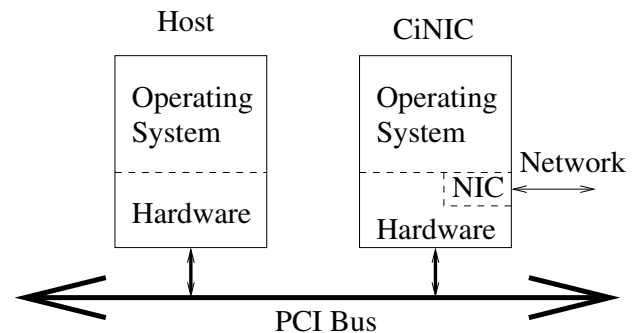
Figure 1. Logical Diagram of CiNIC Architecture

The only major limitation of the current architecture is that its hardware configuration is fixed. Because the demands placed on a network processor may differ significantly from those on a general purpose processor, design decisions based on the expected task-set of one may not be applicable to the other. That is, there might be significant opportunities to improve performance or functionality of the CiNIC if it were possible to tune the hardware architecture.

The next generation architecture for CiNIC intends to incorporate programmable logic into the design in or-

der to explore the benefits of architectural flexibility. One such technology that is well suited to the task is Altera's NIOS processor[1, 2]. This is a 16 or 32 bit microprocessor and peripherals that can be programmed onto a Field Programmable Gate Array (FPGA). This technology provides a processor, peripherals, and custom logic all on one chip. Another reason we chose to evaluate the NIOS processor was the availability of the Microtronix Linux Development Kit[6]. The Microtronix LDK contains hardware expansions for an existing NIOS development board and a port of the uClinux operating system. This provided us an FPGA development board running uClinux that we used to test the performance and features of the NIOS technology.

This paper is an overview of the current CiNIC architecture and discusses our experience evaluating a new programmable logic-based hardware architecture. The first section is a brief overview of the hardware and software architecture of the current CiNIC system. The second section is a discussion of the goals for the next generation CiNIC architecture. The third is brief overview of the NIOS technology and Microtronix uClinux. The final section presents the results of some experiments used to evaluate the capabilities and performance of the NIOS system.

2 Overview of CiNIC Architecture

The goal of the CiNIC architecture is to build a network co-processing engine that is completely transparent to the user. The current CiNIC architecture consists of a coprocessing system mounted in a host computer. The coprocessing system is running a full Linux 2.4 operating system which provides a software development environment, a TCP/IP stack, and services such as a firewall and QoS support. It is this complete Linux implementation that makes our coprocessing scheme different from others. Instead of the host operating system processing a network function, the parameters for that function call are packaged up and shipped to a dedicated coprocessor for execution. The results are propagated back to the host system after the network function is completed[4].

The intent of offloading the entire TCP/IP stack is to save valuable host processing cycles by allowing network functions to be executed in a coprocessor. In addition to processing savings, one can add network functionality such as firewalling, QoS, and traffic shaping without adding a processing load to the host. Similar to the way that video cards began to offload video processing, the CiNIC architecture takes steps to providing offloaded network functions.

The original CiNIC architecture began with just a simple coprocessor board, but evolved when it was realized that running a second operating system within the coprocessing system provided many benefits. The key to running two operating systems within one system is the use of a non-transparent PCI-to-PCI bridge to separate the devices of the host system and the devices of the coprocessor system. Communication between the two operating systems is accomplished via shared memory scheme in which the host maps some of the coprocessor memory into its own address space.

2.1 Hardware Overview

The current CiNIC coprocessor is an Intel EBSA-285 board comprising the following components:

- SA-110 microprocessor running at 233 MHz,

- 21285 core logic chip set,

- 148 MB of system memory for development purposes,

- a set of flash memory for BIOS images,

- and a serial port to use as a console.

Although a real commercial network coprocessor would not need all of these components, they are included to provide for our development environment with the greatest possible flexibility. This coprocessor card is plugged into an Intel 21554 EB backplane[5]. The bridge is a non-transparent PCI-to-PCI bridge that separates the host PCI devices from the co-host PCI devices, effectively creating primary and a secondary PCI bus. The host processor controls devices on the primary bus and the coprocessor controls devices on the secondary bus. Figure 2 is a detailed diagram of this configuration.
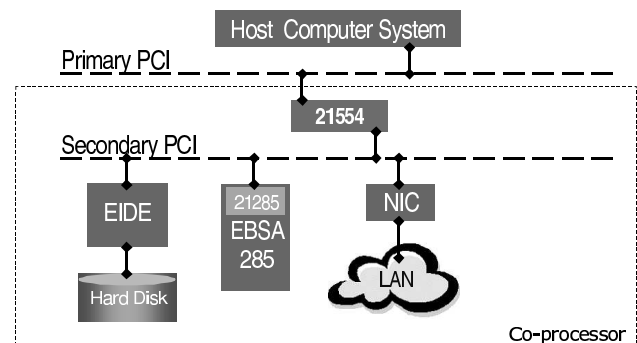
Figure 2. Detailed Diagram of CiNIC Architecture

The addition of a secondary PCI bus, allows any devices having Linux drivers to be easily added to the coprocessing system. In the case of Figure 2, an IDE disk has been added to allow for web caching. What are not shown in Figure 2 are potential devices on the host computer system.

The piece which allows the two computer systems to co-exist is the non-transparent PCI-to-PCI bridge. From both the host computer system and the coprocessor viewpoint, the bridge appears to be a single PCI device, but is responsible for moving data between the two buses. To do this the bridge performs address mappings and address translations. An address range in the primary bus can be

mapped to an address range in the secondary and vice-versa. This allows the bridge to respond to PCI transactions destined for the other side by performing address translation as shown in Figure 3.
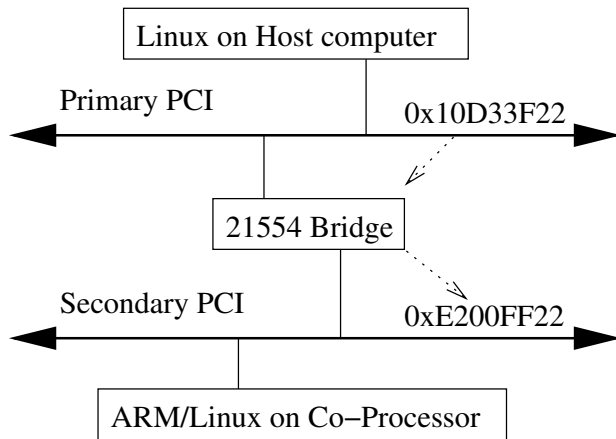
Figure 3. Address mapping through the 21554 bridge

In the example shown in Figure 3, a transaction is started on the primary bus at address 0x10D33F22. The bridge sees this address is in the range of addresses it is translating. Using an internal translation base register, it maps the host address of 0x10D33F22 into the co-processor's address space on the secondary PCI bus at 0xE200FF22. It then carries out the transaction on the secondary PCI bus using this translated address. If this were a read, then the bridge would return the read data on the primary PCI bus. In the case of a write, the bridge would execute the write on the secondary PCI bus using the data passed to it by the primary bus [9].

2.2   Software Overview

The goal of the system software is to make the coprocessor invisible to applications while providing all TCP/IP state processing. The entire data transfer process between the two computing systems is based on a shared memory scheme. This shared memory is located in the coprocessor's memory and is mapped to the PCI space of the secondary bus via registers in the 21285 core logic. This secondary PCI space is in turn, mapped to the primary PCI space via the bridge. It is in this way that the host system can access the SDRAM of the coprocessor operating system.

To execute the host's network functions, the coprocessor must intercept and redirect the TCP/IP stack calls to the network coprocessor. In the CiNIC architecture this is done at the socket layer. When an application makes a socket system call, a kernel module takes over the processing of the system call. It marshals the parameters for the call, places them in shared memory and invokes the coprocessor to complete the function call. When the coprocessor has

completed the function call, it passes the parameters back to the host. The host saves process cycles by scheduling another process while waiting for the socket call to return.

The example above is the ideal operation of each system call. Many situations are much more complex. Functions such as select() which can execute on local file descriptors and sockets, needed to be coded with care. Once the functions were coded, the end result is a library of socket system calls that to the applications and user behave exactly like the standard system calls. Standard applications such as FTP were used to test the functionality of this coprocessing architecture[4].

3   Goals for the Next Generation CiNIC Architecture

The current CiNIC architecture is based on a commercially available microprocessor. Our goal is to move to an architecture using programmable logic which will make more parts of the system customizable. While there is nothing fundamentally wrong with the old CiNIC architecture, a completely customizable system will allow us to experiment with the hardware aspects of offloaded network functions and give us the ability to adapt the hardware architecture to match the needs of the software. This new architecture will no longer be static, but can be changed to enhance the functionality and performance of the coprocessing system. The new architecture that was tested is based on Altera's Excalibur technology: RISC processors synthesized and running on programmable logic. The processor provides the power to run an operating system and the flexibility of programmable logic to implement custom hardware on the chip. It is a classic example of hardware/software co-design: designing hardware and software concurrently to meet system-level objectives[3].

This architecture allows us to conveniently decide what to implement in hardware and what to implement in software. With the added capability of programmable logic, much more of the system is under our control. It would allow us to use a physical shared memory, rather than just mapping the SDRAM of the coprocessor to the host. Since we utilize very few of the PCI bridge's capabilities, this is another component that could be replaced with a more elegant, custom solution. The original CiNIC architecture was developed to aid in network performance analysis and experimentation. Moving the entire platform to programmable logic will provide extra features and allow further analysis and experimentation.

4   NIOS and Microtronix uClinux

The core piece of this project is the Altera NIOS development board. The Altera board is part of a technology they call Excalibur. It is based on running RISC microprocessors on a FPGA and providing room for user-defined logic. The NIOS technology is a soft-core processor, mean-

251

ing that the hardware for the processor is implemented in a Hardware Description Language. This code is compiled and programmed into the FPGA, making part of the FPGA into a microprocessor. One advantage of this architecture is the ability to use leftover logic on the FPGA for processor peripherals. This allows peripherals like a UART, Ethernet controller, PCI Logic, or IDE controller to be implemented on the same chip as the microprocessor. In addition to peripherals, the architecture lends itself to easy implementation of user-defined logic on the chip. This allows the functionality of a microprocessor, peripherals, and ASIC all on the same chip (see Figure 4). The result is an architecture with extremely flexible software and hardware, perfect for embedded systems work[1].
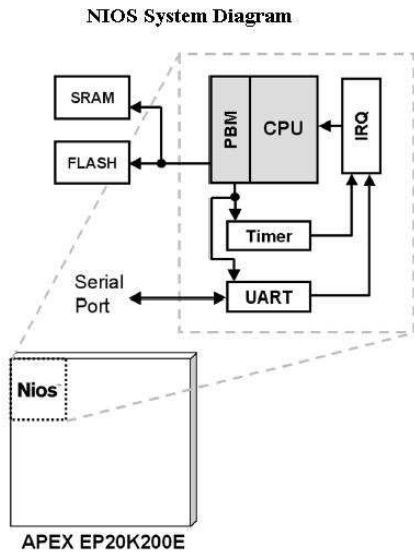

Figure 4. NIOS System Diagram[2]

The Microtronix Linux Development Kit (LDK) is a hardware and software expansion to the NIOS board. It contains extra memory, in the form of 16 Mbytes of SDRAM, a real time CMOS clock, and an 10BaseT Ethernet controller chip plus transceiver. The software expansion is a port of uClinux, a real-time, embedded version of Linux. The embedded aspect refers to the small memory footprint of the operating system kernel. Also ported were the standard Linux tools and commands, cp, ls, etc. With the addition of this kit, the NIOS processor had a fully functional operating system. For more information on the contents of the Microtronix LDK see [6].

5 NIOS System Evaluation

The evaluation of this technology to use as a next generation architecture comprises two aspects: performance and capabilities. First, we needed to determine whether the FPGA processing system could handle the Linux network processing code. Next experimentation needed to be done

in order to determine how easy it is to add custom logic to the system and how that custom logic integrates with the Linux operating system. Once these two aspects were analyzed, it was shown that the NIOS system was a good target for moving the next generation CiNIC architecture.

5.1 Performance

Since the purpose of the CiNIC architecture is to improve performance related to network processing, we conducted a series of timing measurements to determine network processing latencies on the new coprocessor architecture. The experiment consisted of running a test application under uClinux on the NIOS development board and measuring the time required to traverse the stack. The test application sent marked packets using the send() function. A timer was added to the hardware of the FPGA specifically for counting how long a send() function took to execute on the NIOS system. The server that was receiving the packets simply discarded them, since the timing was all done while the data traversed the TCP/IP stack of the NIOS system.
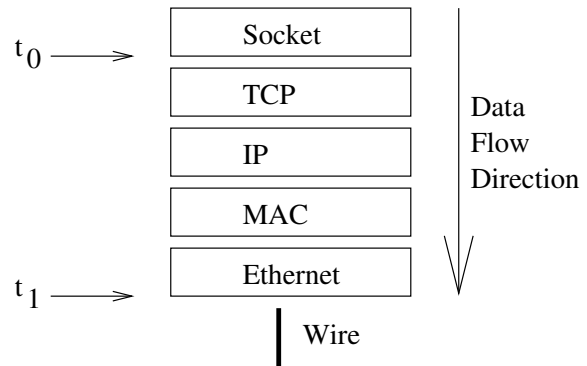
Figure 5. Model of uClinux TCP/IP stack

The timing model is relatively straight-forward. The timings for each measurement were based on the buffer size of the send() function. These buffer sizes range from 64 bytes to 64,000 bytes. The timer for the system is started when the send() function enters kernel space. The timer is stopped after all the data for the send has been written to Ethernet controller. Figure 5 shows the approximate placement of the timing hooks in the stack.

The hooks themselves are snippets placed into the kernel stack that start, read, and stop the timer. In order to make the ends of the buffers recognizable, each buffer was marked with 6 0xAA characters at the front and 6 0xEE characters at the end. On each buffer size, 2 timestamps were recorded: entry and exit times. *Entry time* is defined to be the time when the first packet of the buffer (the one with the 0xAA's) hits the t_0 marker (see Figure 5). The *exit time* is defined to be the time when the last packet of the buffer (the one with the 0xEE's) hits the t_1 marker.

This test was run 25 times for each buffer size. From there we took the median value of the 25 runs and calcu-

Adjusted NIOS TCP/IP Stack Timing
25 trials per buffer size
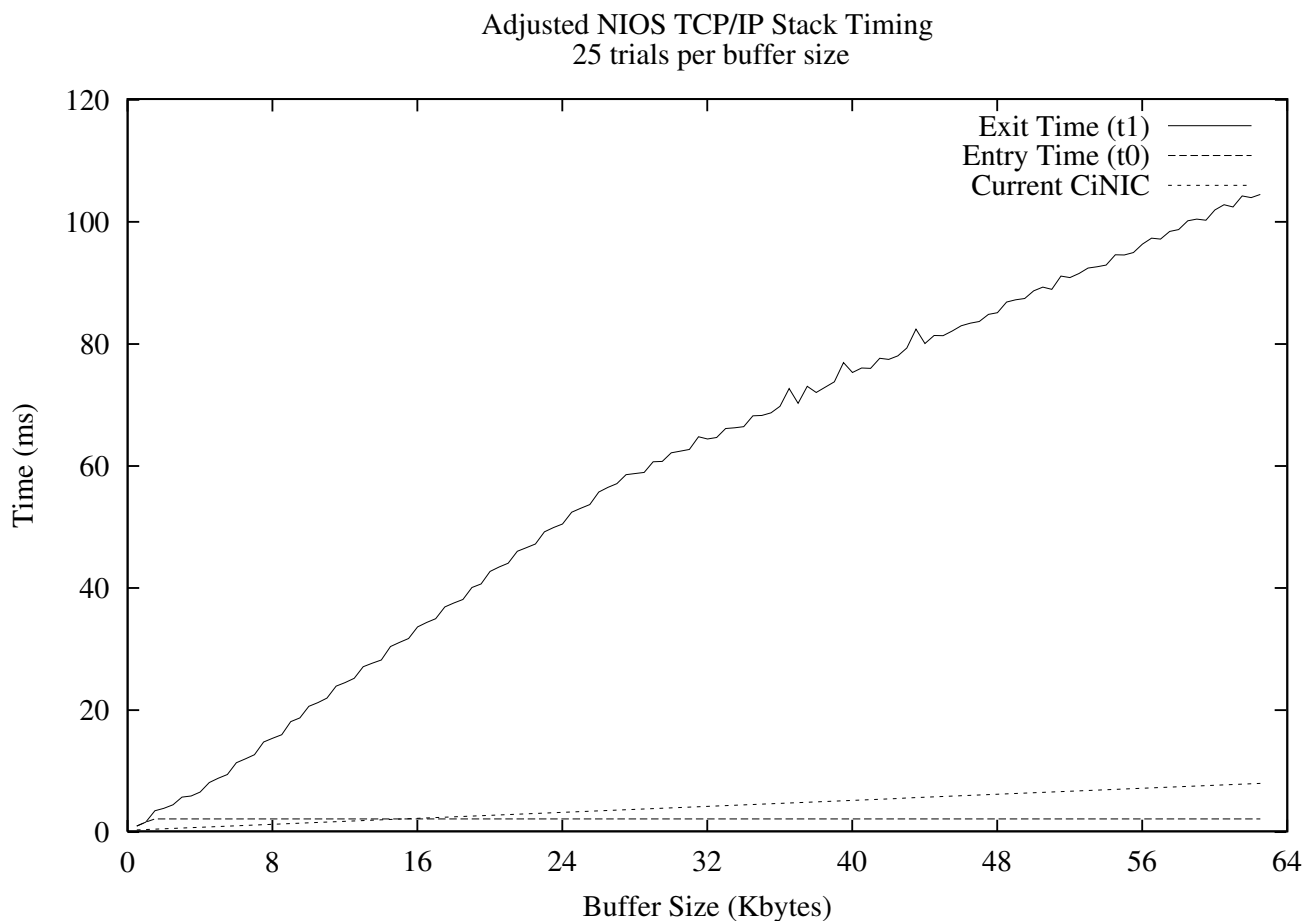
Figure 6. Adjusted NIOS TCP/IP Stack Timing

lated the standard deviation of each of the measurements. The timings showed that 99.5% (3 standard deviations) of the measurements were within 3% of the median value. This means that the uClinux stack on the NIOS system has a consistent processing time. Figure 6 shows the results—adjusted to eliminate the cost of the timestamping itself—and compares them to the results of the current CiNIC TCP/IP stack timings.

The entry time stayed constant as was expected. Whether sending a buffer of 64,000 bytes, or 1500 bytes, the time it takes for the first packet to hit the Ethernet controller is the same. The exit time is proportional to the buffer size that was sent. That is, the processing time is linearly dependent on how much data needs to be sent.

The TCP/IP stack measurements for the NIOS system give us an idea of the network performance of uClinux running on a NIOS processor. When comparing this performance measurement to the measurements from the CiNIC system[4], the NIOS processing appears to be approximately an order of magnitude slower. The reasoning for this, however, can be attributed to three major differences in the hardware and software between the two systems.

First, the architectures and processor speeds of the

two systems are different. The current CiNIC using a StrongARM processor runs at 233MHz while the NIOS processor runs at 33.333 MHz. That is a difference of 7x. Second, there is a difference in the Linux kernel versions. The CiNIC's kernel version was 2.4 and the NIOS has version 2.0. There are large differences in the TCP/IP stack portion of these two kernels. Many of the modifications between the 2.2 and 2.4 kernels were improvements to network processing. Further analysis in this area is needed, but when comparing these results it is important to remember that the NIOS and CiNIC are running similar (but different) stacks.

The third difference is the fact that the NIOS has a 10 BaseT Ethernet interface, while the tests for the CiNIC were done with a 100 BaseT Ethernet interface. This issue suggests a difference of 10x if the network interface is the limiting factor. This hypothesis was easily tested since the driver for the controller chip contained a "chip busy" state. There is a finite amount of memory for packets in the chip, so when this is full the driver will wait and try again. A single debug statement proved that the driver was entering the "chip busy" state during multiple packet sends. This shows the NIOS processor and the uClinux stack are capable of

generating packets faster than 10 Mbps. The addition of 100 Mbps Ethernet would allow us to determine if the processor is capable of 100Mbps speeds.

5.2 Capabilities

This section discusses the use of programmable logic within the NIOS FPGA. The idea of adding custom hardware is one of the main advantages of this programmable system. The process of modifying or adding hardware can be done quickly, giving us flexibility not present in the current CiNIC architecture. The process of examining the capabilities of custom hardware, was implementation of a very simple custom hardware device and integrating it with uClinux.

The simple hardware device was a two register device which used control values in register A to decide when to perform a one's complement on the value in register B. This was implemented in Verilog code, along with the interfacing logic necessary for tapping into the NIOS system bus. From the NIOS system side, a user-defined interface was setup on the bus and assigned a region in the memory-mapped space. This allowed all software to see the registers of the custom hardware at a particular physical address. Since NIOS does not contain an MMU, it only has to dereference a pointer to that physical address to access the registers.

The process of adding custom hardware to the system was found to be very painless. Once the initial busing structure was understood. Integration with uClinux was also simple due to the lack of a virtual addressing for the NIOS processor. Our custom hardware wrote data to register B, flipped a bit in register A, waited for the hardware to flip the A bit back, and read out the "processed" value. This allowed us to implement a simple scenario, where the processor waits for the custom hardware to finish[1].

6 Conclusion

This paper presented a network coprocessing architecture based on an FPGA running a modified version of the Linux operating system. The architecture will allow us to implement changes and analyze the hardware/software trade-off very quickly. We studied both the performance and capabilities of this architecture. In the performance aspect, the NIOS under-performed the Strong Arm based CiNIC. But, it was shown that this may have been due to the NIOS's 10 Mbps Ethernet versus the CiNIC's 100 Mbps Ethernet. We feel confident that the NIOS, FPGA based architecture can adequately perform network functions. Also, if Moore's law holds true, the performance of the FPGA based architecture will continue to improve rapidly, providing more performance in the near future. In the capa-

bilities arena, NIOS was found to be very flexible and fit our needs from an ease of use and development standpoint. The uClinux operating system provided us with the needed software support and the ability to add custom hardware to the system.

References

[1] ALTERA CORP. *Excalibur Backgrounder, Version 1.0*, 1.0 ed., June 2000.

[2] ALTERA CORP. *NIOS Embedded Processor—Getting Started, Version 1.1*, Mar. 2001.

[3] DE MICHELI, G., AND GUPTA, R. Hardware/software co-design. *Proc. IEEE 85*, 3 (Mar. 1997), 349–365.

[4] FISCHER, J. CiNIC—CalPoly intelligent NIC. Master's thesis, California Polytechnic State University, June 2001.

[5] INTEL CORP. *21554 PCI-to-PCI Non-Transparent Bridge Evaluation Board User's Guide, Version 1.0*, Jan. 2001.

[6] MICROTRONIX CORP. *LDK Getting Started Guide, Version 1.5*, Sept. 2001.

[7] ROSU, M., SCHWAN, K., AND FUJIMOTO, R. Supporting parallel applications on clusters of workstations: the intelligent network interface approach. In *Proceedings of the Sixth IEEE International Symposium on High Performance Distributed Computing* (Portland, OR, Aug. 1997), pp. 159–168.

[8] STEENKISTE, P. Analysis of the Nectar commuinication processor. In *IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems* (Feb. 1992), pp. 1–3.

---

[1]NIOS 2.0 has the capability to provide up to five custom instructions. I.e., there now is a direct implementation for custom hardware instructions.

Viable Remote Sensing Multispectral / Hyperspectral Bandwidth Compression with Region of Interest Prioritization Capability

John Saghri

Electrical Engineering Department
Cal Poly State University

Final Report

Project 55410

Viable Remote Sensing Multispectral / Hyperspectral Bandwidth Compression with Region of Interest Prioritization Capability

John Saghri
Electrical Engineering Dept., Cal Poly, San Luis Obispo, CA 93407
jsaghri@calpoly.edu

December 2003

Summary

The major tasks of the project "Viable Remote Sensing Multispectral / Hyperspectral Bandwidth Compression with Region of Interest Prioritization Capability," were completed successfully. The results of this project are described in the attached professional papers "KLT/JPEG 2000 Multispectral Bandwidth compression With Region of Interest Prioritization Capability," which was presented and published at the SPIE International Symposium in San Diego, August 2003. Preliminary results, based on statistical, and machine-based exploitation measures, show significant improvement in the compression performance as a result of incorporating the region-of-interest capability in our baseline compression algorithm. We successfully applied our developed compression system to both multispectral and hyperspectral imagery data. For hyperspectral data, due to their inherent high spectral dimensionality, we are also considering a two stage spectral decorelation process that potentially achieves better results. Preliminary results, as described at the end of this report, are promising.

1

KLT/JPEG 2000 Multispectral Bandwidth compression with Region of Interest Prioritization Capability[1]

John A. Saghri*[a], Andrew G. Tescher*[b], Anthony M Planinac*[c]

[a]Electrical Engineering Dept., Cal Poly, San Luis Obispo, CA. 93407, jsaghri@calpoly.edu
[b]AGT Associates, 14670 Fieldstone Drive, Saratoga, CA 95070, andytescher@attbi.com
[c]Electrical Engineering Dept., Cal Poly, San Luis Obispo, CA. 93407, tplaninac1@charter.net

ABSTRACT

The region of interest (ROI) coding feature of JPEG 2000 image compression standard is extended to multispectral imagery. This is accomplished by enabling ROI capability of JPEG 2000 module in the previously developed Karhunen-Loeve/JPEG 2000 compression of multispectral images. Preliminary results, based on subjective, statistical, and machine-based exploitation measures, show significant improvement in the compression performance. Depending on the ROI/background relative size and the desired quality differential, the improvement in the classification accuracy can increase by as much as one hundred percent without an increase in the bandwidth.

Keywords: Bandwidth compression, JPEG-2000, Multispectral image compression, Region of Interest coding, Classification, Thematic map, Confusion matrix.

1. Introduction

The region of interest (ROI) coding is a desirable feature in a viable image bandwidth compression system. This feature allows the limited bandwidth to be distributed optimally over the various regions of the scene based on the relative importance of the features they contain. In the past, we have investigated several approaches to address this issue[1, 2, 3]. In the latest attempt[3], we used a dual compression approach to introduce a feature prioritization capability to our previously developed Karhunen-Loeve Transform/ Discrete Cosine Transform (KLT / DCT) multispectral bandwidth compression system[4,5]. The dual system included; (1) a primary unit for conventional coding of multispectral image, and (2) an auxiliary unit to code the resulting compression-induced error incurring on features, or regions, of interest. The regions of interest were determined via a spectral unmixing procedure which transformed the original set of multispectral images into another set of images where each image showed the concentration of one, and only, feature within the geographical area being imaged. Via computer simulation, we demonstrated that this scheme did provide a feature discrimination capability in the compression system. Features of interest in the scene maintained a relatively higher fidelity (precision level) in the compression process than the nonessential features. However, the added computation complexity (which translates directly to additional weight, size, and power requirement) due to the secondary compression unit and the spectral unmixing procedure makes this approach challenging for onboard implementation. For this reason we are proposing to replace the above-mentioned feature prioritization scheme with a simpler and more viable scheme based on the recently announced ROI capability of the standard JPEG-2000[6,7]. Leveraging on the fully developed standard technology reduces the cost and complexity of this approach. The regions of interest can be selected manually by the user or they can be set automatically using a thematic map obtained via an unsupervised classification procedure[5].

The JPEG 2000 compatible ROI schemes include the General Scaling and Maxshift methods[6,7]. The ROI shape is limited to a rectangle or ellipse. The general scaling allows the ROI/background quality differential to be

---

2

adjusted freely by the user at the encoder. However, the decoder must generate the ROI mask. In the Maxshift method, the decoder need not generate the mask. However, the user has no control over the ROI/background quality differential. The differential depends directly on the quantization step size and the dynamic range of the original image.

To assess the improvement in performance resulting from of application of ROI, we used both statistical and machine-based exploitation measures. For the statistical measure we used the root mean square versus the coding bit rate. For the machine-based exploitation measure, we used unsupervised classification process. Preliminary results indicate the inclusion of ROI coding significantly improves the performance.

2. Compression Block Diagram

Figure 1 shows the overall compression block diagram. The ROI can be selected manually by the user. Alternatively,   as depicted in the diagram, ROI they can be set using a thematic map generated from an unsupervised classification procedure[5]. The process begins with a set of multispectral images. The set typically contains five to sixteen images, each having a specific spectral band. The image set is first transformed via the KL Transform to generate of set of spectrally uncorrelated principal component images. The principal component images are then individually coded by JPEG 2000 with the ROI feature enabled. At the receiver, the process is reversed. The principal component images are first reconstructed via JPEG 2000 decoder. Note that the principal component images will have a higher resolution, i.e., reconstruction accuracy, in the regions of interest than the background region. The set of decoded principal component images are transformed via inverse KLT to obtain the reconstructed images.

3. Performance Evaluation Methodology

Three different schemes are selected to assess the effectiveness of KLT/JPEG 2000 multispectral compression with ROI utility. These methods include subjective, statistical, and machine-based exploitation measures. Subjective measure is performed by visual inspection only. For statistical measure, the root mean square of the compression-induced error is used. For assessing impact of compression on machine-based exploitation, an unsupervised classification procedure[9] was implemented. The procedure results in a thematic map which depicts various classes, i.e., features, in the scene being imaged. To quantify the accuracy of the resulting thematic maps, a confusion matrix is used. A confusion matrix is an efficient and convenient method to assess the impact of compression-induced error on the classification. It compares the relation between known reference data (ground truth) and the result of the automatic classification, so it tells how well the training samples of each class have been classified. The confusion matrix columns represent ground data (assumed to be correct), and the rows represent map data (classified by the automatic procedure). The diagonal elements represent agreement between ground and map, so ideally the matrix should have all zero off-diagonal elements.

The *errors of omission* (map producer's accuracy) are the incorrect in column divided by total in column, hence measures how well the map maker was able to represent the ground features, i.e. it indicates how well training set pixels of the given cover type are classified.

$$\text{Producer's Accuracy} \; = \; \frac{\text{Number of correctly classified pixels per category}}{\text{Number of reference pixels used for that category}}$$

$$= \; \frac{\text{Diagonal element}}{\text{corresponding column total}}$$

3

The *errors of commission* (map user's accuracy) are incorrect in row divided by total in row, hence measures how likely the map user is to encounter correct information while using the map, i.e. indicates the probability that a pixel classified into a category actually represents that category.

$$\text{User's Accuracy} \ = \ \frac{\text{Number of correctly classified pixels per category}}{\text{Number of pixels classified in that category}}$$

$$= \ \frac{\text{Diagonal element}}{\text{corresponding row total}}$$

The map *overall accuracy* is the total on diagonal divided by the grand total.

$$\text{Overall Accuracy} \ = \ \frac{\text{Total number of correctly classified pixels}}{\text{Total number of reference pixels}}$$

$$= \ \frac{\text{Sum of diagonal elements}}{\text{Sum of columns (or rows)}}$$

For our experiments, the pixels in the original thematic map are treated as being correctly classified, i.e., the ground truth. Thus, any difference between two thematic maps, i.e., one obtained from the original and the other from the compressed/reconstructed image sets, represents misclassified pixels, or classification error. The confusion matrix can efficiently capture the classification errors. Figure 2 shows the methodology to evaluate the classification accuracy.

4. Experimental results

For JPEG 2000, we used the JJ2000 Java-based software package from Ecole Polytechnique Federale de Lausanne (EPFL). This software package is a product of the EPFL-Canon-Ericsson university/industry joint team efforts to create a software package to meet the JPEG2000 standard requirements. We opted to use the Maxshift method for ROI feature of JPEG 2000 as it was the simplest to use and did not require the decoder to generate the ROI masks, i.e., it did not involve any overhead information to be transmitted to the decoder to identify the ROI

The test imagery was from LANDSAT 5 TM space-borne sensor, and covered a rural/urban scene in Montana in seven unequal bands in the spectral range of 0.45-2.35 micrometers. There were a total of seven spectral bands. Each spectral band was composed of 512x512 pixels with each pixel having a footprint of 30 meters (with the exception of band 6 with a 120m footprint). This test set was selected because it contained a diverse range of natural terrain and, as such, very challenging for classification. Figure 3 shows spectral band 5 of the seven-band Montana test set showing the rectangular 220x 237 ROI. The image contrast is enhanced for display purpose. For this experiment, a single rectangular ROI was chosen over a farmland. Selection of ROI based on thematic map will be implemented in the near future. Figure 4 shows the ROI-enabled reconstructed image in figure 3 at 0.34 bit per pixel (bpp) overall coding bit rate. The higher reconstruction quality of ROI is with respect to its background region is clearly shown. The overall coding bit rate of 0.34 bpp was obtained by averaging the individual coding bit rates of the seven principal component images. The coding bit rate for individual spectral bands were selected so as to produce uniform quality in the reconstructed spectral bands[8]. Figure 5 shows the reconstruction of ROI section of spectral band five at 0.34 overall coding bit rate with and without ROI capability.

4

Figure 6 shows the overall coding bit rate versus the averaged root mean square error, with and without ROI feature. The plot shows that the advantage of ROI coding diminishes at the higher coding bit rates. This is expected because at a high coding bit rate, the ROI/background quality differential approaches zero. Figure 7 depicts the thematic maps of ROI associated with the original image set and 0.34 bpp reconstructions with and without ROI capability. Each of the three thematic maps shows the distribution of five different classes within the geographical area being imaged. Each class represents a species with a unique spectral signature. Table 1 shows the confusion matrix representing the differences between the thematic maps of the ROI associated with the original image set and 0.34 bpp reconstructions *without* ROI capability. Table 2 shows the confusion matrix representing the differences between the thematic maps of the ROI associated with the original image set and 0.34 bpp reconstructions *with* ROI capability. At this rate, application of ROI utility causes the overall classification accuracy to increase by nearly 100 percent.

5. Conclusion

It was shown that the region of interest coding feature of JPEG 2000 image compression standard can be extended to multispectral imagery as well. This was accomplished by enabling ROI capability of JPEG 2000 module in the previously developed Karhunen-Loeve/JPEG 2000 compression of multispectral images. Preliminary results, based on subjective, statistical, and machine-based exploitation measures, showed significant improvement in the compression performance. Depending on the ROI/background relative size and the desired quality differential, the improvement in the classification accuracy can increase by as much as one hundred percent without an increase in the bandwidth. Further work need to be carried out to evaluate the effectiveness of various available ROI coding schemes as well as ROI selection schemes.

ACKNOWLEDGEMENT

REFERENCES

1. J. A. Saghri, A. G. Tescher, and M.J Omran ",Feature-Based Lossy Compression of Multispectral Data", IEEE 1999 International Geoscience and Remote Sensing Symposium, Vol. 6/99, July 1999
2. J. A. Saghri, A. G. Tescher, and A. Boujarwah ",Spectral-Signature-Preserving Compression of Multispectral Data," Journal of Optical Engineering, vol.38, No. 2, December 1999
3. J. A. Saghri, A. G. Tescher, and M.J Omran ",Class Prioritized Compression of Multispectral Imagery Data", Journal of Electronic Imaging, Vol. 11, No. 2, April 2002
4. A. Saghri, A. G. Tescher, and J. T. Reagan ",Practical Transform Coding of Multispectral Imagery", IEEE Signal Processing Magazine, Vol.12, No.1, January 1995
5. A. Saghri, A. G. Tescher, and F. E. Kozak ",Analysis of of JPEG versus JPEG-2000 foe KLT-based Compression of Multispectral Imagery Data", Proceedings of the SPIE, Vol. 4790, No. 33, July 2002
6. R. Grosbois, D. Santa-Crus, and T. Ebrahimi ",New Approach to JPEG-2000 Compliant Region of Interest Codinga," Proceedings of the SPIE 46th Annual Meeting, Applications of Digital Image Processing XXIV, San Diego, July 2001
7. C. Christopoulos, J. Askelof, and M Larsson ",Efficient Methods for Encoding Regions of Interest in the Upcoming JPEG2000 Still Image Coding Standard" Stockholm, IEEE Signal Processing Letters, Volume 7, Number 9, pp.247-249, September 2000.
8. J. A. Saghri, A. G. Tescher, and F. E. Kozak, ",Analysis of of JPEG versus JPEG-2000 foe KLT-based Compression of Multispectral Imagery Data", Proceedings of the SPIE, Vol. 4790, No. 33, July 2002

5

Original Imagery

ISODATA Classificatio

Thematic Map showing Region-of-interest

Karhunen-Loeve Transformation

Compression

Spectrally-Uncorrelated Imagery

Standard *JPEG 2000*
(With Region-of-interest capability)

Transmission Channel

Inverse* JPEG *2000

Reconstruction

Inverse Karhunen-Loeve Transformation

Reconstructed Imagery
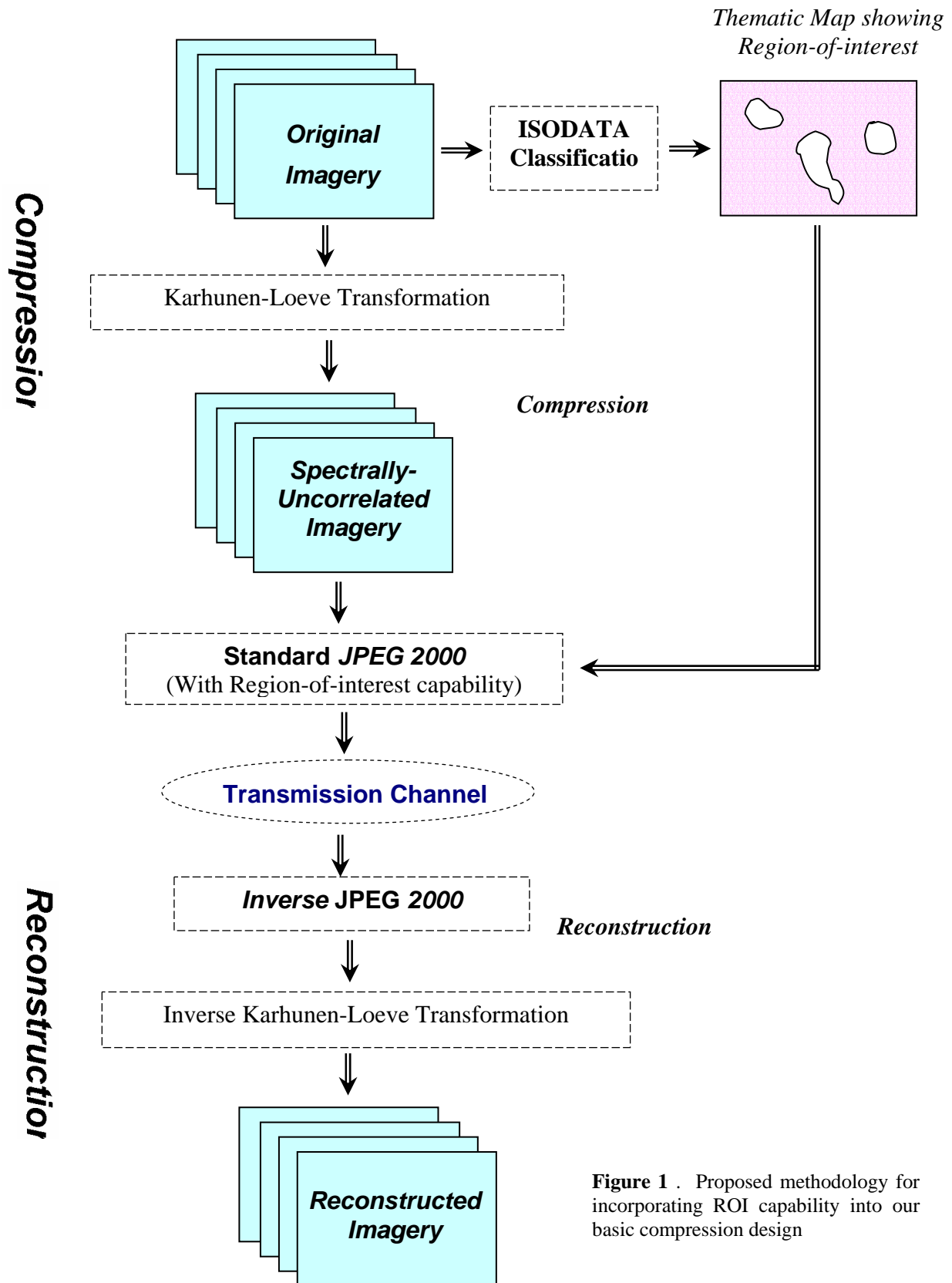
Figure 1 . Proposed methodology for incorporating ROI capability into our basic compression design

6
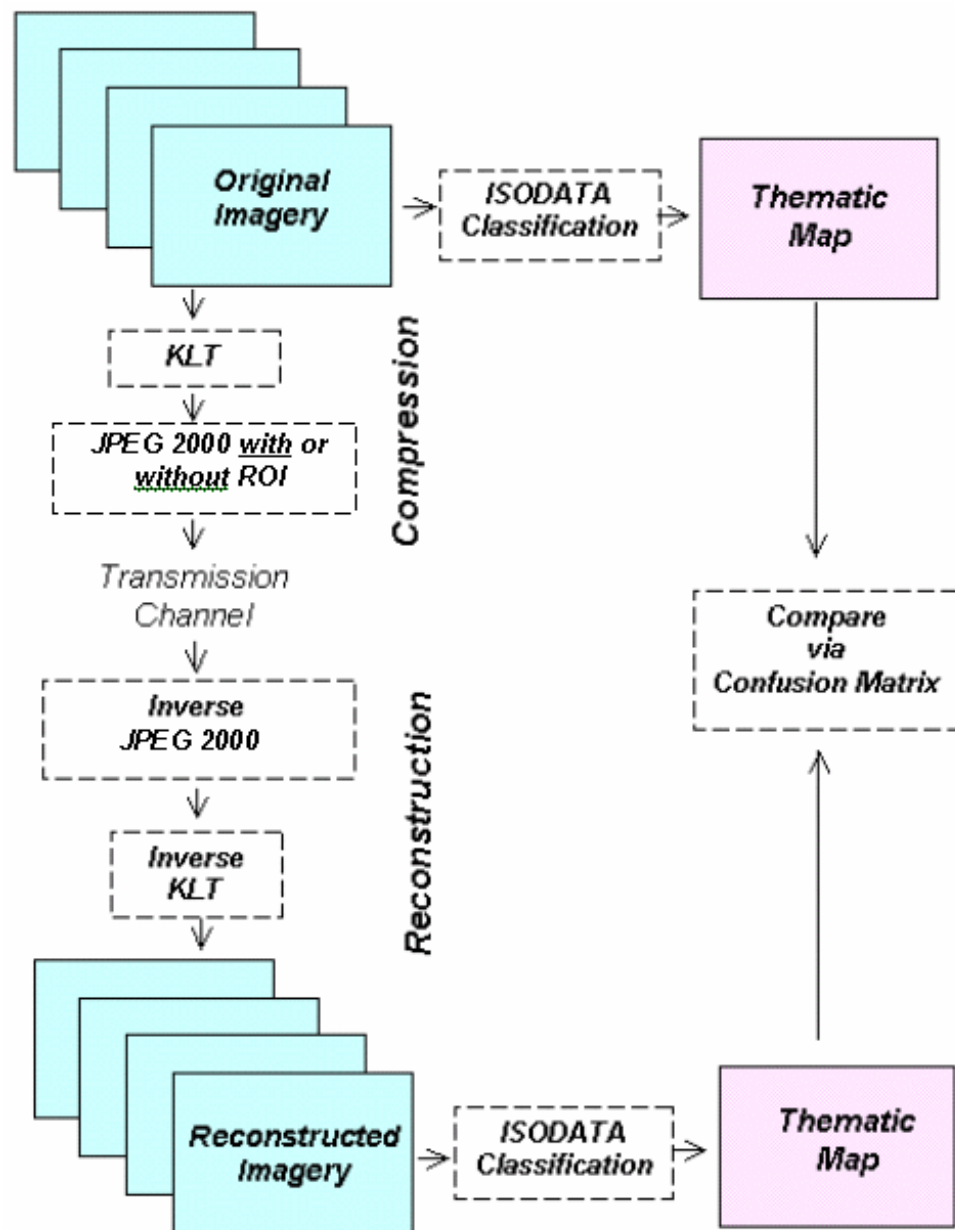
Figure 2. Methodology to evaluate the impact of compression on classification

7
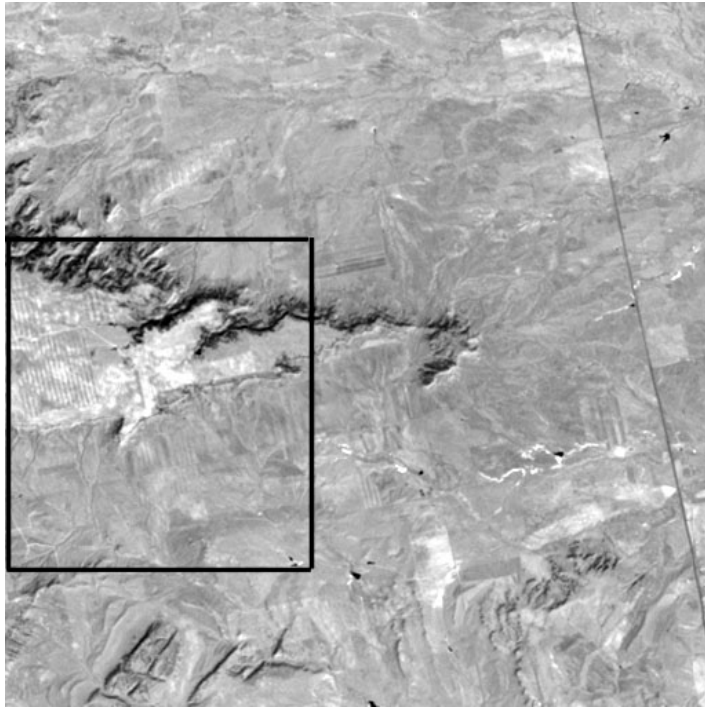
Figure 3.   Band 5 of Montana 7-band test set (LANDSAT) showing the selected rectangular ROI
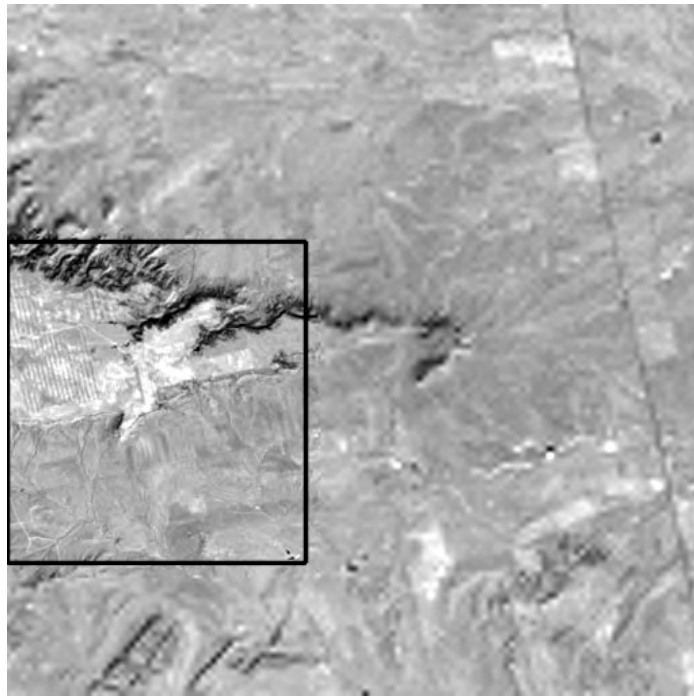
Figure 4.   Reconstructed Band 5 of the total seven bands at an overall coding rate of 0.343 bpp with ROI

8
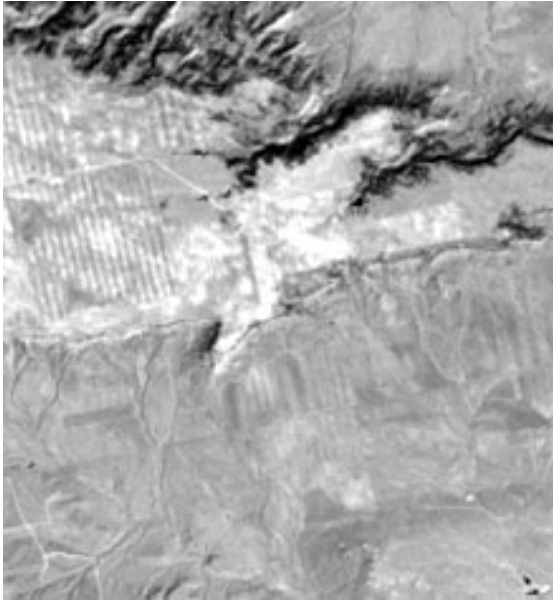
(a) Original Band 5 ROI
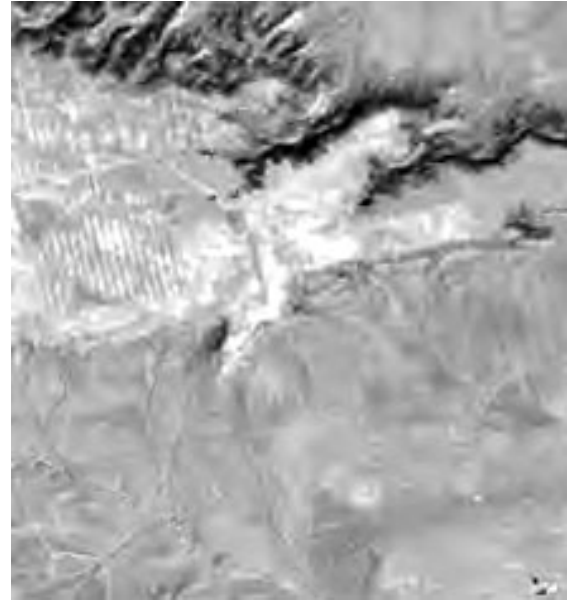
(b) Reconstruction *without* ROI at 0.343 bpp
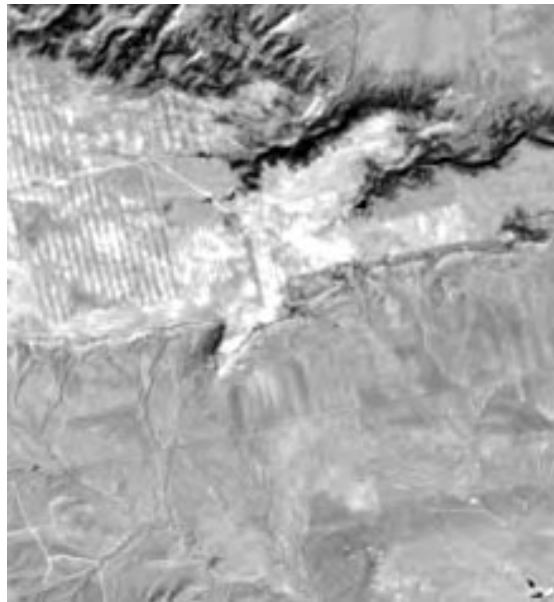
(c) Reconstruction *with* ROI at 0.343 bpp

Figure 5. Comparioson of ROI reconstruction quality at an overall coding bit rate of 0.34 bpp . (a) original 220x237, (b) 0.34 reconstruction without ROI capability, (c)).34 bpp reconstruction with ROI capability

9

Rate versus Error

Figure 6. Overall coding bit rate versus averaged root mean square error with and without ROI feature.
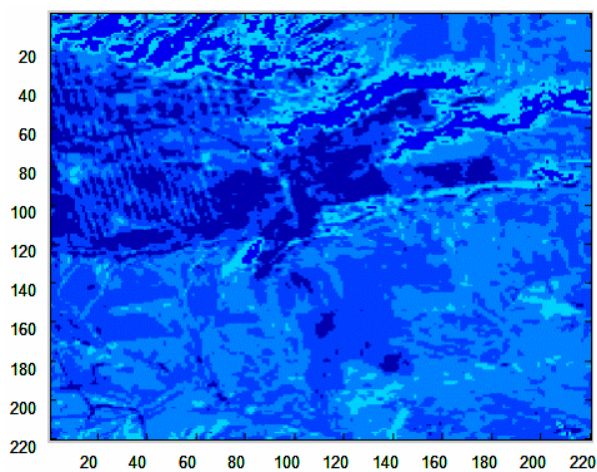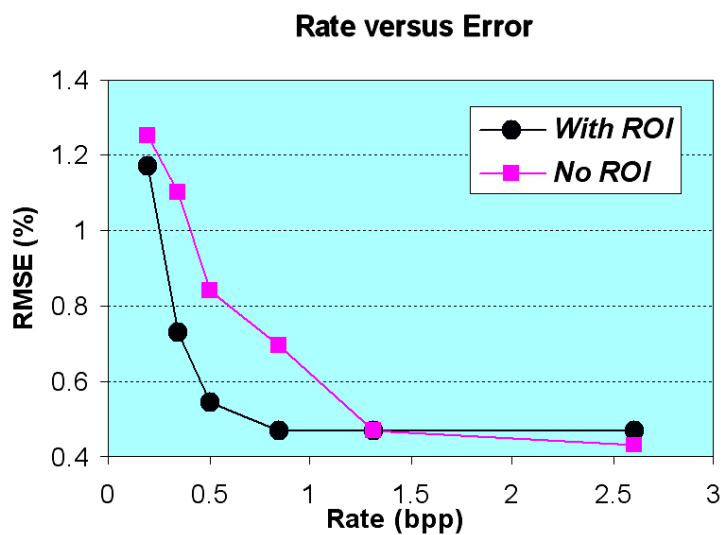
(a) Original Thematic Map
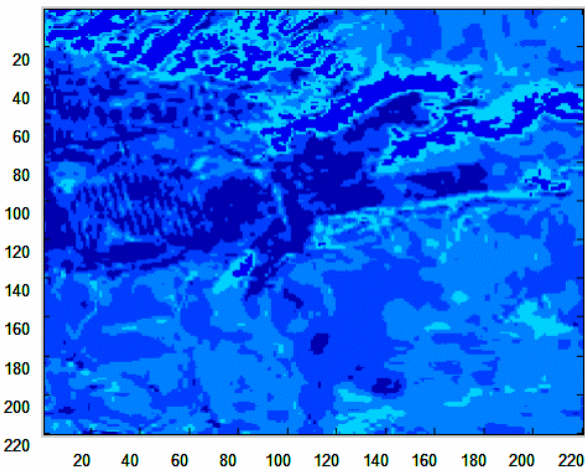
(b) Thematic Map *without* **ROI** Capability

(c) Thematic Map *with* **ROI** Capability
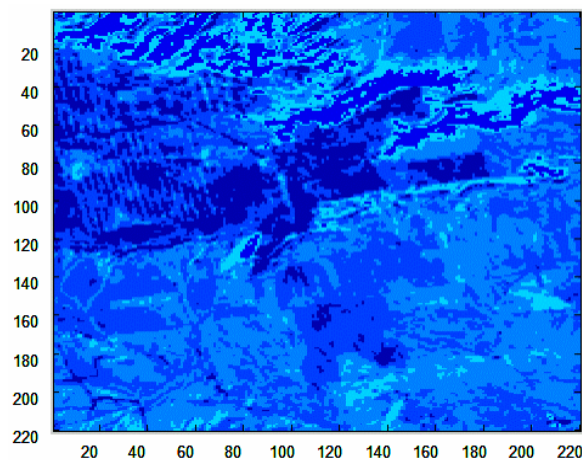
Figure 7. Thematic maps of the ROI with a five classes , **(a)** original, (b) obtained from reconstruction at 0.34 bpp without ROI capability, and (c) obtained from reconstruction at 0.34 bpp with ROI capability

10

Without ROI (Rate = 0.34 bpp)

Confusion matrix

	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	2624	99	2368	780	208
Class 2	50	1699	165	204	997
Class 3	2404	244	7691	5441	734
Class 4	883	390	5252	9404	1559
Class 5	323	830	1046	1565	1440

Overall accuracy =   0.47 %

Table 1. Confusion matrix showing the changes in the classification (compared to classification based on the original) at 0.34 bpp coding rate *without* ROI utilization.

With ROI (Rate = 0.34 bpp)

Confusion matrix

	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	5623	0	456	0	0
Class 2	0	3014	0	0	101
Class 3	394	0	14809	1311	0
Class 4	0	0	1345	15671	472
Class 5	0	86	0	572	4546

Overall accuracy =   0.90 %

Table 2. Confusion matrix showing the changes in the classification (compared to classification based on the original) at 0.34 bpp coding rate *with* ROI utilization.

11

Hyperspectral Bandwidth Compression

We successfully applied our developed compression system to hyperspectrul imagery data as well. Due to the inherent high spectral dimensionality of hyperspectral data, we are also considering a two stage spectral decorelation process that potentially achieves better results. This process is briefly described below.

For our testing, we used hyperspectral data obtained from a HIRIS (Hyperspectral InfraRed Imaging System) instrument onboard airplane. The data consists of a total of 192 spectral bands each encompassing 256 by 256 pixel mosaic corresponding to the same geographical terrain on earth. The original dynamic range of 12 bits has been linearly mapped to 8 bits.

Our baseline scheme was to divide the set of 192 spectral bands into contiguous sets of 16 spectral images and then coding each set independently using the same method we used for the multispectral images (KLT followed by JPEG 2000). This method is a simple and straightforward and fully exploits the spectral redundancy between the bands of each set. However, it does not exploit the spectral redundancy between sets of spectral images. Theoretically, the optimum method is to consider the entire 192 spectral images as belonging to a single set and applying Karhunen-Loeve transformation (KLT) to the entire set to achieve full spectral decorelation. The required KLT operation would involve obtaining the eigenvectors and eigenvalues of a 192 by 192 covariance matrix; a process which is not practically feasible due to mathematical complexity and computation time. An alternative solution is as follows:

1) Divide the 192 spectral images into nine sets of 16
2) Apply KLT to each set of 16 spectral images to obtain 16 principle component images per set
3) Place each column of 16 principle component images from each set next to the 16 principle component images from the adjacent set.
4) The outcome will be a matrix of principle component images consisting of 16 columns and 9 rows.
5) Apply KLT again to each of the first three rows of the resulting matrix of principle components

The above cross KLT operation theoretically achieves the same full spectral decorelation achievable by doing KLT on the entire 192 spectral bands in one step (not computationally feasible).

We apply the cross KLT to the first three rows of the principle component matrix. This is because, the first three principle components require over 90 percent of the required coding bandwidth in the subsequent JPEG 2000 (with ROI capability) compression process. Hence it pays off to further reduce the coding bit rate for these principle components via further

exploiting their spectral redundancy (cross KLT operation). The lower principle component images require very little bandwidth to code. Hence, for these principle component images, the realized gain in the coding bit rate is too small to offset the added computational complexity of the additional cross KLT operations.

The preliminary results so far are very promising. The figure below show the bit rate versus error (percent root mean square error) for the HIRIS data using the both approaches (with and without cross KLT). The graph shows that at 0.33 percent rms error, the coding bit rate is reduced from 0.58 to 0.32 bits per pixel. This translates to a saving of about 50 percent in the bit rate as a result of applying cross KL operation. We are still working on optimizing this alternative method. In particular, we are changing the bit allocation strategy for JPEG 2000 coding of the principle component images to achieve better results. More, and hopefully better, results along with detailed discussion of this method, will appear in the master thesis dissertation of Cal Poly graduate student Tony Planinac who is expected to present and defend his thesis in February 2004.
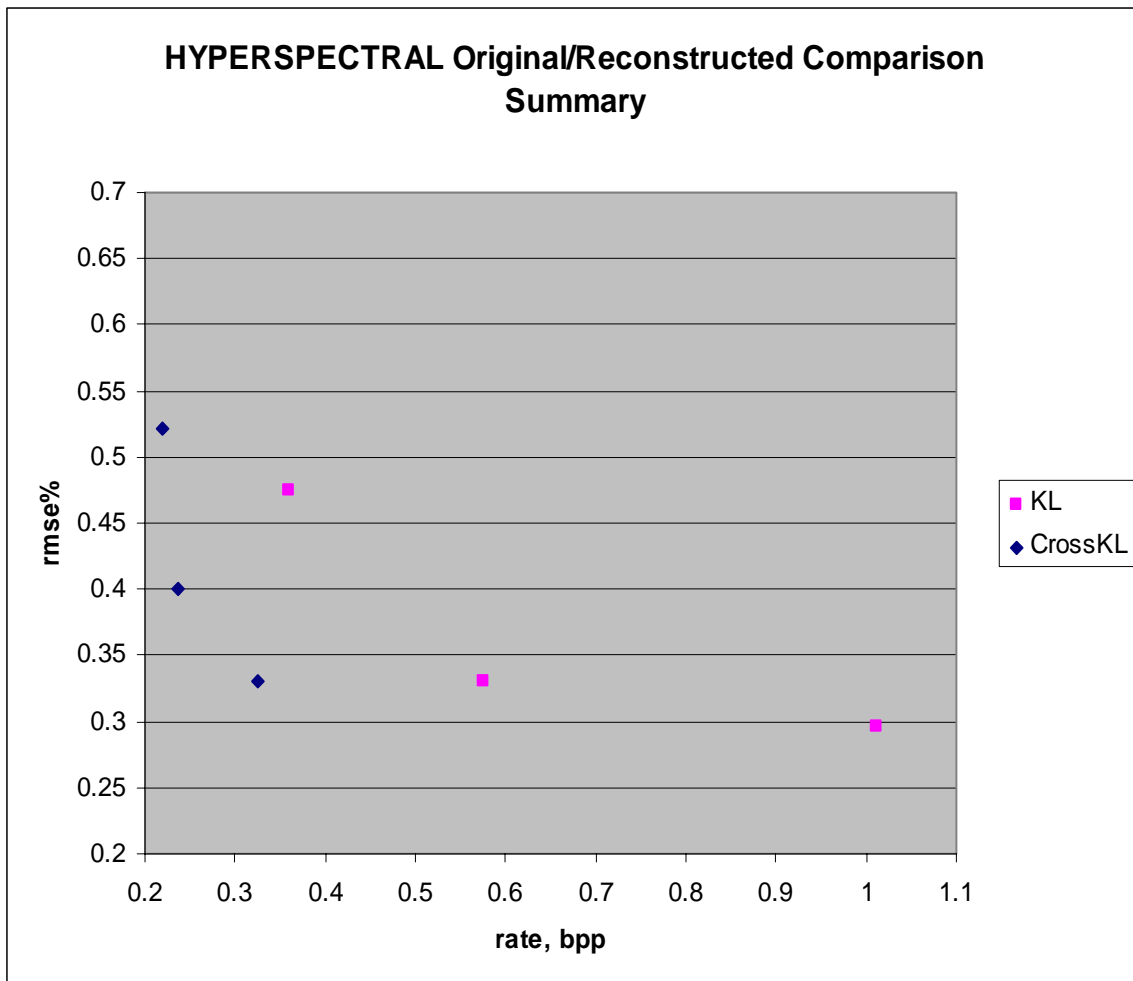
Construction of Tetrahedral Photonic Bandgap Crystal: Demonstrating Three-Dimensional Self-Assembly using DNA Linkage

Peter V. Schwartz

Physics Department
Cal Poly State University

Project name: Construction of Tetrahedral Photonic Bandgap Crystal: Demonstrating Three-Dimensional Self-Assembly using DNA Linkage

Principle Investigator: Peter V. Schwartz, *Cal Poly Physics Department*

This project is truly multidisciplinary, involving faculty on Cal Poly campus from physics, chemistry, biology, and material science, and is done in collaboration with the UCSB Materials Engineering and Chemical Engineering Departments.

The $20,000 grant was spent almost entirely on salaries. Four students worked full time for the entire summer, and another student worked half time. Additionally, this grant paid half the summer salary for the PI. Most of the expensive materials were purchased by our collaborators at UCSB, so little of our budget was required for materials ($500).

Technical Background:
Short strands of DNA (10–20 bases) are used as nanoscopic, selective Velcro® to guide the self-assembly of nanostructures in solution because DNA strands only bond (hybridize) to other DNA strands if they have a complementary sequence. A three-component system is used in our studies (see Figure 1); where a linker is hybridized to two separate surface-bound DNA strands in order to achieve aggregation of microspheres.
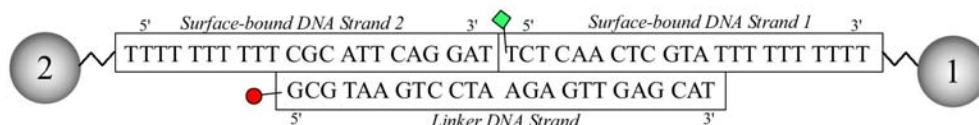
Figure 1. A schematic of the three-component system used in our studies. The green square and the red cycle represent the fluorescence labels used to quantify the DNA coverage and hybridization efficieny described in the "Progress" section. Many thousands of identical DNA strands are on each microsphere.

We have successfully achieved DNA attachment to single polystyrene microspheres and subsequent controllable aggregation of these microspheres by means of DNA hybridization (Figure 2).
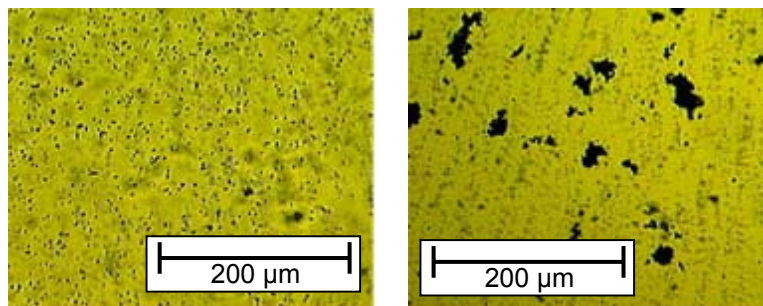
Figure 2. Micrographs of aggregation of microspheres via DNA hybridization. Left: no linker present results in no aggregation. Right: linker present results in aggregation.

Progress:
Quantifying DNA surface coverage. In more recent work, we are quantifying the DNA coverage (amount of DNA bonded to the microsphere), and the subsequent hybridization efficiency (the percentage of bonded DNA that attach to freely-floating DNA strands via hybridization) through measurements using fluorescent-labeled DNA. The concentration of DNA in solution is also measured using light absorption techniques in order to confirm measurements. Figure 3 and Figure 4 show the schemes we have used to quantify the DNA coverage and the hybridization efficiency, respectively. Quantification of the DNA coverage and hybridization efficiency has allowed us to optimize the procedure we use to aggregate microspheres. We

1

have found that the surface density of DNA bonded to the beads can be controlled by the concentration of DNA in solution and the pH of the reaction. We have also found that the hybridization efficiency decreases at higher DNA surface coverage. We have also designed and constructed an improved observation cuvette to do fluorescence spectroscopy on our opaque suspensions of microspheres (Figure 5), and plan on repeating our measurements with improved accuracy.

Figure 3. Scheme to measure DNA coverage.

Figure 4. Scheme to measure linker DNA hybridization on surface-bound DNA.
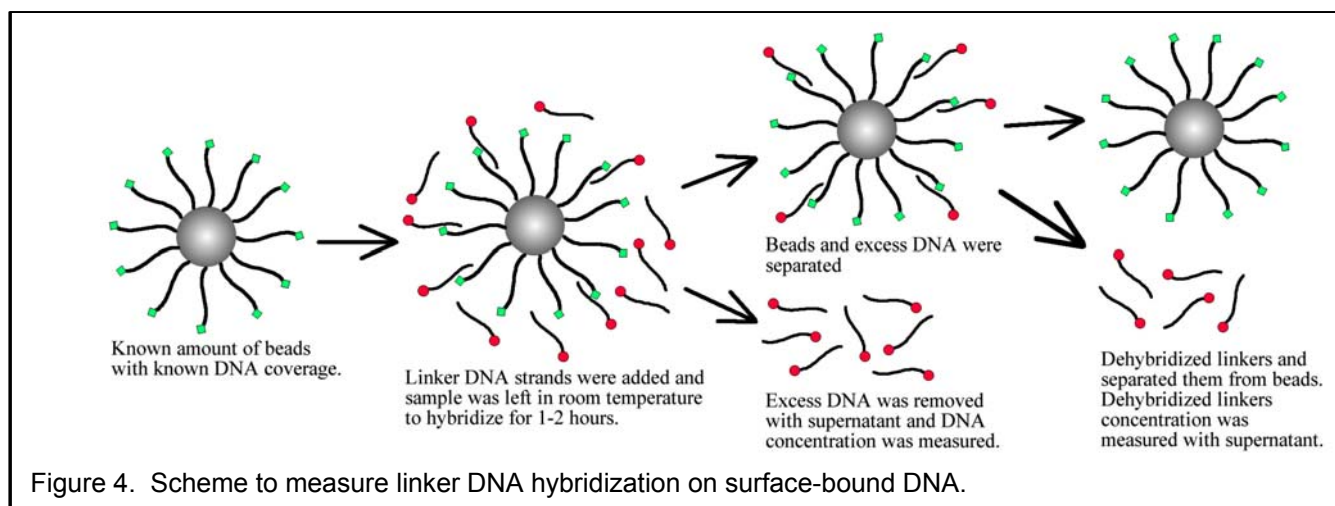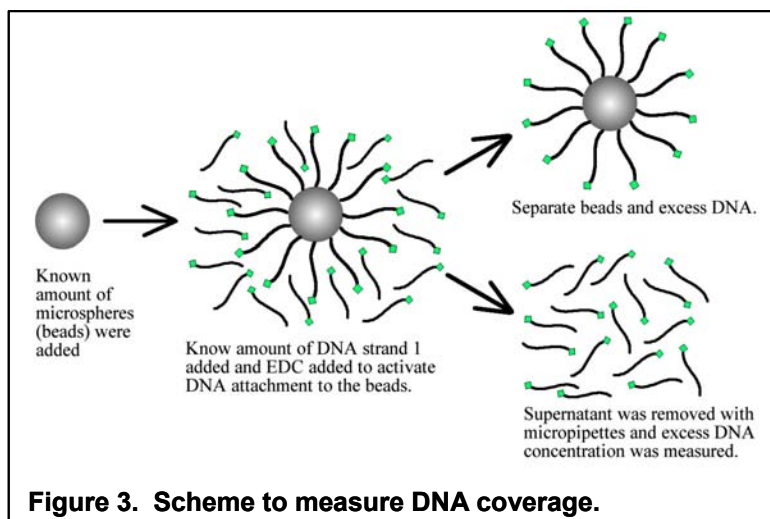
Improving Specificity of Bonding
Although our preliminary results are very encouraging, we are taking an additional step to investigate how to produce pure aggregates and how to anneal out imperfections and impurities. We have introduced into the experiment an impurity in the form of "random strand beads" – microspheres covered with a DNA sequence that is not complementary to anything else in suspension. The random strand beads are also a different color, so that they can be identified. Our efforts are focused on eliminating random strand beads from the aggregates made from hybridized beads. We have discovered that lowering salt concentration and providing gentle agitation during hybridization can improve the size and purity of the aggregates. A large aggregate made of yellow microspheres is shown in Figure 6. The random strand beads are white and are largely absent from the aggregate. Using fluorescence microscopy, we can identify the minority impurities inside the aggregate. We have designed and built a heating stage that fits on optical and fluorescence microscopes, and have begun preliminary annealing experiments (Figure 7). We will observe aggregates during future annealing experiments and hope to be able to expel impurity beads through heating.

Figure 5. Cuvette designed for fluorescence measurments on opaque samples.

2

Investigating a new Substance

An aggregate of DNA-linked microspheres is a new substance with unique properties. We have been able to make macroscopic samples of microsphere aggregate and future experiments will quantify the visco-elastic properties as well as the dependence on temperature.
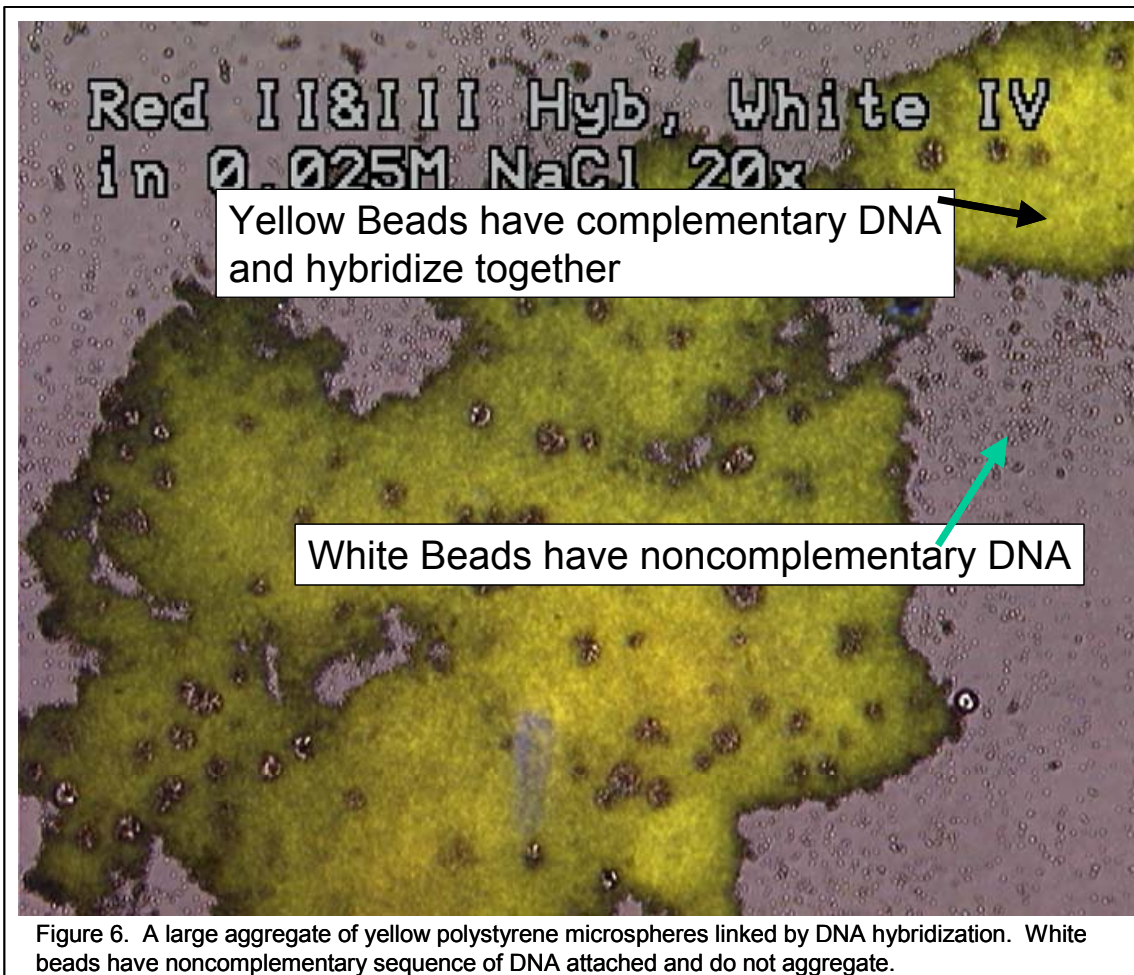
Figure 6. A large aggregate of yellow polystyrene microspheres linked by DNA hybridization. White beads have noncomplementary sequence of DNA attached and do not aggregate.
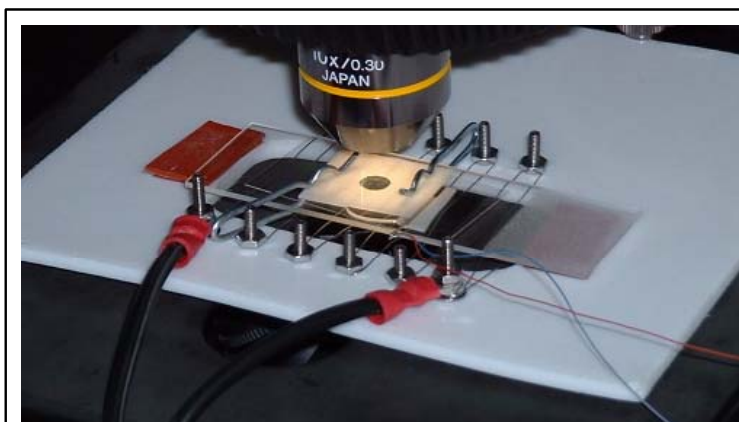
Figure 7. Heating stage to anneal and melt aggregates while under optical and fluorescence microscopy

3

Development of Launch Interfaces for the CubeSat Program

Clark Turner
Computer Science Department

Jordi Puig-Suari
Aerospace Engineering Department

Cal Poly State University

Development of Launch Interfaces for the CubeSat Program
Final Report

Led by Stanford University and Cal Poly, the CubeSat program represents an effort to address some of these issues by developing a new class of standard pico-satellites.

C3RP Funding provided support for the following CubeSat related activities at Cal Poly.

- Qualification of the standard Cubesat deployer (P-POD) for space flight: Cal Poly students completed the design and testing of the P-POD (see Figure 1) which is now qualified for space flight. This effort included the development of deployment electronics compatible with standard launch vehicle deployment signals. The P-POD serves two main functions. 1) The P-POD is a standard housing for the CubeSat satellites; allowing universities to focus on manufacturing their satellite and not worry about integration with the launch vehicle. 2) The P-POD protects the primary payload and the launch vehicle from the university satellites in the event of a failure.
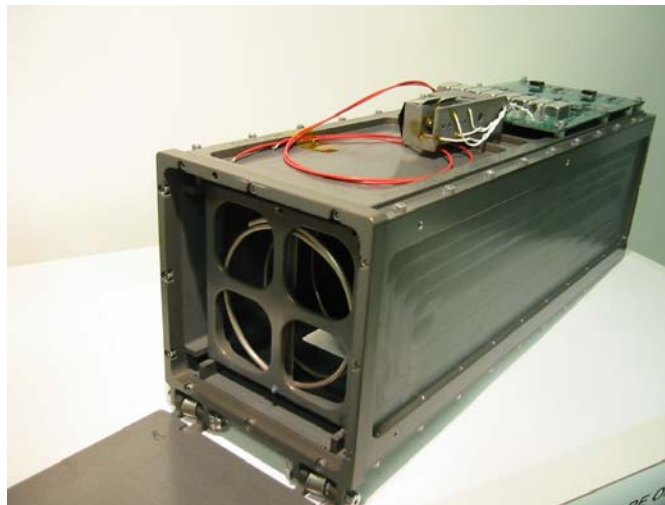
Fugure 1. CubeSat Standard Deployer

- Support of first CubeSat launch: The first CubeSat launch took place in June 30, 2003 on a Russian Rockot launch vehicle (see Figure 2). This launch was coordinated by the University of Toronto. Cal Poly provided two P-POD deployers for this launch. The two P-PODs successfully deployed the following satellites.
    - o P-POD 1:
        - CANX-1 (University of Toronto)
        - DTUSat (Technical University of Denmark)
        - AAUSat (Aalborg University, Denmark)

    - o P-POD 2:
        - Quakefinder, (QuakeSat, Palo Alto, CA)
        - Note: QuakeSat was a triple CubeSat that took the entire volume of one P-POD.
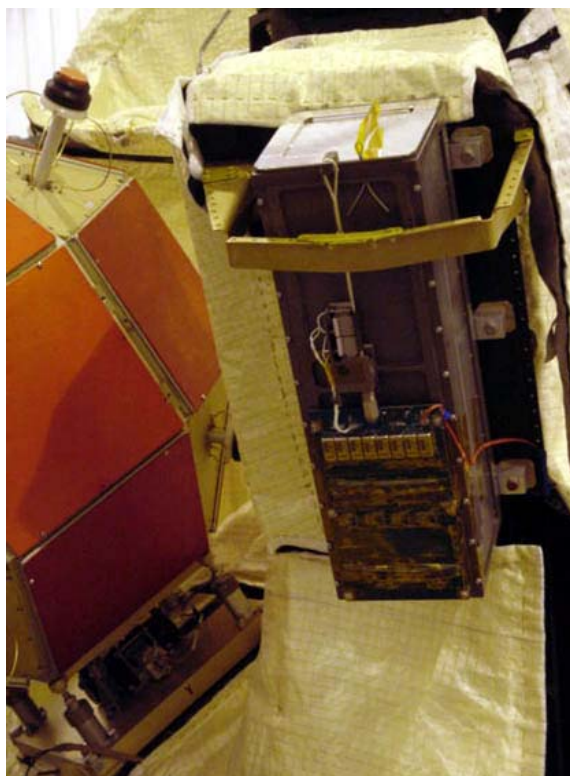
Figure 2. P-POD attached to the Rockot vehicle and ready to go.

The experiences from this launch were presented in the following paper at the 17th AIAA/USU Conference on Small Satellites in August 2003:

"Affordable, Low-Risk Approach to Launching Research Spacecraft as Tertiary Payloads,"
F. Pranajaya, R. E. Zee-Space Flight Laboratory, University of Toronto Institute for Aerospace Studies; P. Lundahl Thomsen-Danish Space Research Institute; R. Wisniewski-Aalborg University; M. Blanke-Danish Technical University; L. Franklin-QuakeFinder, LLC; J. Puig-Suari-California Polytechnic State University

- Coordination of a Second CubeSat launch in the second quarter of 2004: The Cal Poly CubeSat team is completing negotiations with Kosmotras, the company marketing the Dnepr launch vehicle, to provide a second CubeSat launch in the fall of 2004. That launch will consist of 5 P-PODs with 14 university satellites. The universities that are participating are:

US Universities:	**International Universities:**
California Polytechnic State University	Hankuk Aviation University, South Korea
Cornell University	National Cheng Kung University
Kansas State University	Nihon University, Japan
Montana State University	Norwegian University of Science Technology
University of Arizona	
University of Hawaii	
University of Illinois	

Cal Poly's responsibilities for this launch include obtaining export licenses for U.S. satellites, final integration and testing of CubeSats and P-PODs, integrating the P-POD to the launch vehicle, and overall coordination of the launch activities.

- Development of U.S. CubeSat launch capabilities: The Cal Poly team has been working with U.S. launch providers to develop a domestic CubeSat launch capabilities. Launch vehicles under consideration include the Space Shuttle (see Figure 3) and SpaceX's Falcon. A $98K grant from the Department of Defense was recently awarded to Cal Poly to continue development of U.S. launch capabilities for CubeSats.
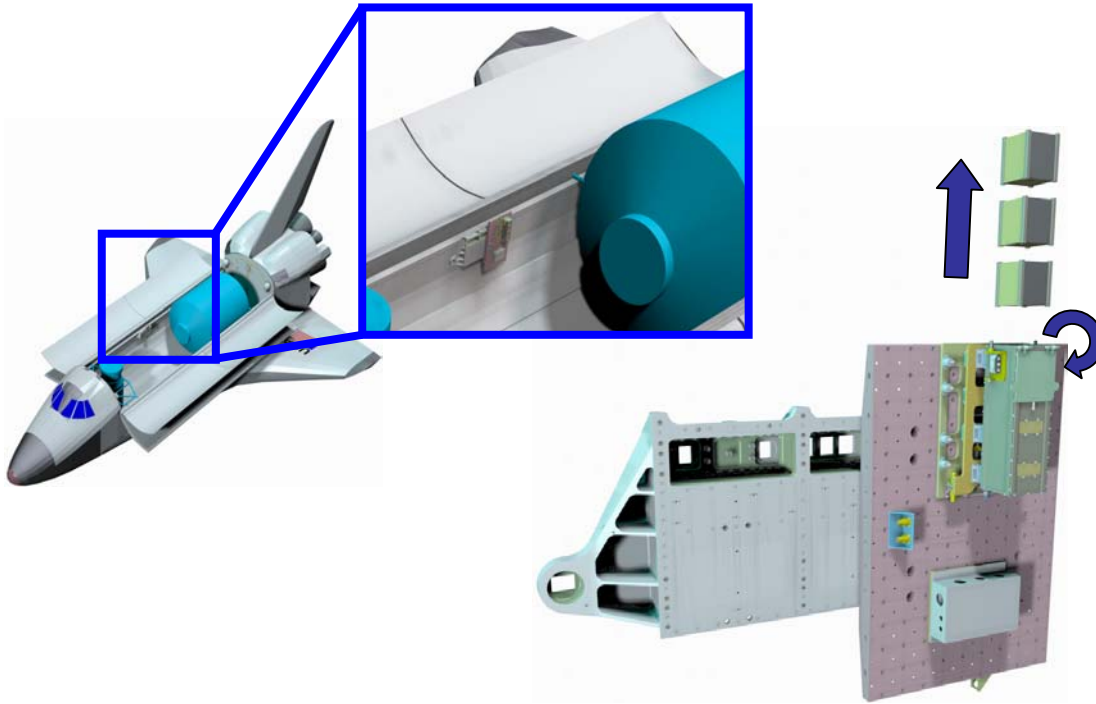
Figure 3 Conceptual Design of P-POD integration to the Space Shuttle developed by NASA

Development of Field Rechargeable Gas Mask Filters

Katherine Chen

Materials Engineering Department
Cal Poly State University

C3RP FINAL REPORT

Development of Field Rechargeable Gas Mask Filters

K. Chen
Materials Engineering, Cal Poly
December 2003

The continued progress of a collaborative project between Cal Poly and Sun Microsystems, Advanced Development was enabled through a partial grant, and has resulted in the creation of a viable testing laboratory for rigorous scientific adsorption and desorption studies for the development of rechargeable gas mask filters. Based on previous results, numerous refinements to prior experimental designs and apparatus were made. Testing capabilities are now possible for future studies to assist in the development of the rechargeable filters.

Project Summary

While the concept of producing a field rechargeable gas mask filter is sound and based on existing technology developed by Dennis Pfister of Sun Microsystems, the ability to rigorously study and optimize the phenomenon is not so straightforward. The ability to control and monitor the adsorption/desorption processes involved with the technology has not been attempted before and proved to be very challenging. Several meetings with the collaborators took place, and several redesigns of the test chamber and the test equipment were made.

In order to increase the signal response during the experiments (i.e., weight gain due to the adsorption of gases), a different adsorbate (R-134) was obtained, and an entirely new testing chamber was designed. The decision was made to create a testing chamber for experiments, while in parallel, construct a working prototype for demonstration purposes. The new test chamber is much more robust in terms of gas leakage and accommodated much more adsorbate and adsorbent. Different adsorbent materials, such as activated graphite, carbon cloths, and carbon foam can easily be exchanged within the new test chamber.

In addition, significant modifications to the experimental set up were made. Laboratory space was dedicated to the project (Figure 1). A more controlled environment was established, and a variable power source was purchased. LabView was used to control the inlet of gas and to acquire real-time data on the chamber temperature, power, and weight. The schematic of the experimental configuration on LabView is provided in Figure 2. The simultaneous acquisition of temperature and weight gain/loss as a function of time enables the scientific understanding of the adsorption/desorption phenomenon.

1

The experimental design will also assist in the optimization of the technology for the gas mask filter. The end goal is to efficiently regenerate adsorption sites of the filter with as little power (and thus minimal temperature gain/loss) as possible.

The combination of people on this project from diverse backgrounds and with different knowledge contributed greatly to the advancement of the project. The objectives of the project that were met are highlighted in Table I. Continued work between Cal Poly and Sun Microsystems is expected to further the development of field rechargeable gas mask filters.

Table I. Objectives of the project met
√ strengthened working relationships between Cal Poly and Sun Microsystems
√ redesigned and built new testing chamber for adsorption experiments
√ acquired essential equipment for testing and reconfigured testing apparatus
√ constructed real-time data acquisition for experiments
√ set up new testing lab for continued studies
√ provided research opportunity to Cal Poly students and staff
√ provided technical expertise to Sun Microsystems Advanced Development
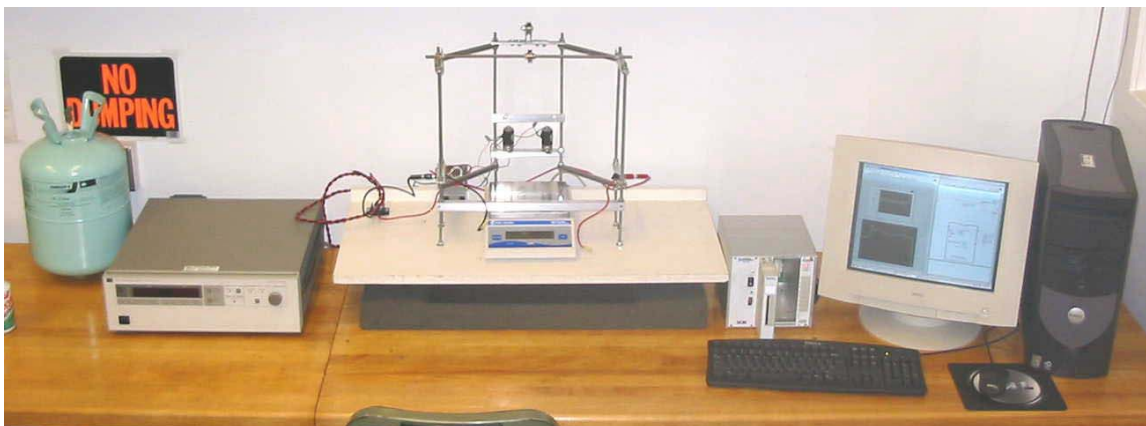
Figure 1. Laboratory set up for adsorption/desorption studies.

2
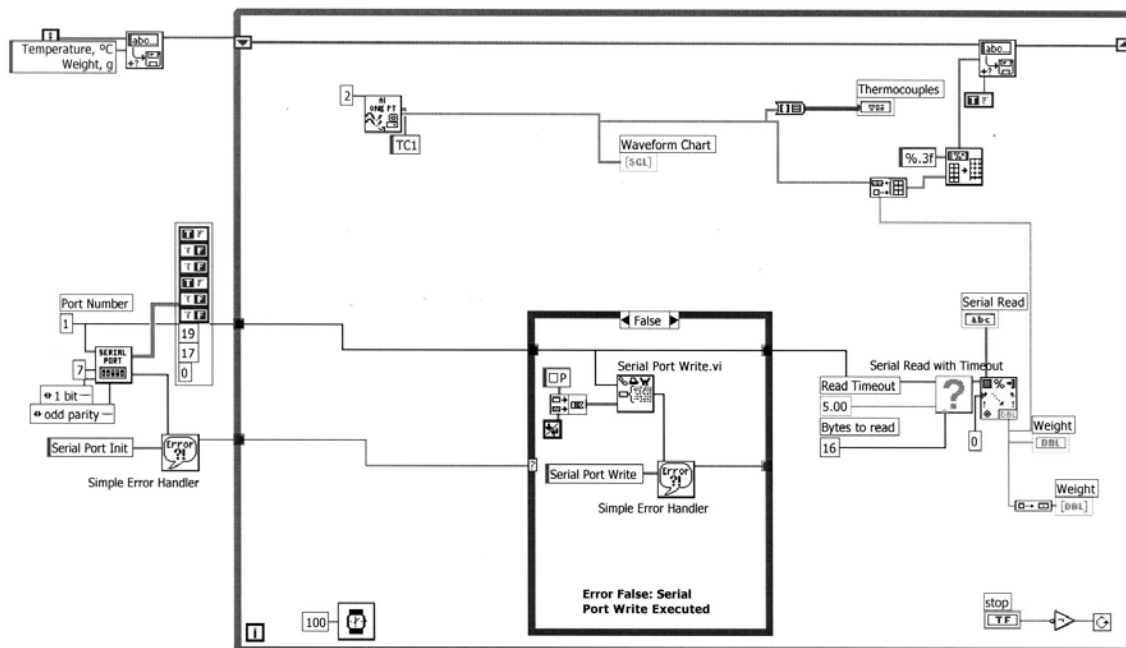
Figure 2. Schematic of the experimental set up on LabView.

A Homogenization Approach to the
Analysis and Design of Composite Systems

Eric P. Kasper
Civil and Environmental Engineering

Joseph D. Mello
Mechanical Engineering Department

Cal Poly State University

A Homogenization Approach to the Analysis and Design of Composite Systems

Eric P. Kasper & Joseph D. Mello

September 29, 2003

Contents

Section 1

Homogenization – Theoretical and Numerical Developments

In this section the notation and methodology used for the development of a homogenization theory after [1] in the setting of multi-dimensional problems with linear constitution and kinematics is presented. Additionally, numerical implementation is also discussed and several numerical examples are explored to show the robustness of the method.

1.1 Theoretical Developments

The asymptotic homogenization theory presented allows for the development of an explicit relation between the microscopic (local) and the macroscopic (global) field variables. The method presented can be summarized into several simple steps:

1. Develop a microscopic weak form for the local representative periodic unit cell. Introduce a set of compatible interpolation functions into the weak form to yield an approximate set of algebraic equations in terms of the unknown characteristic quantities, χ. Given the cell geometry, material response, and periodic boundary conditions solve for the unknown characteristic quantities, χ.

2. Develop a homogenized set of material moduli, $\mathbb{C}^H$, for use in the development of a solution for the global problem, based on the characteristic nodal quantities, χ, and unit cell material moduli, $\mathbb{C}$.

3. Develop a macroscopic weak form for the global continuum problem. Introduce a set of compatible interpolation functions into the weak form to yield an approximate set of algebraic equations in terms of the unknown quantities, $\boldsymbol{u}$ (*i.e.* displacements for

a structural application). Given the global geometry, homogenized material response, boundary and loading conditions solve for the unknown quantities, $\boldsymbol{u}$.

4. Construct either the global or local stress/strain response based on the computed solutions, $\boldsymbol{\chi}$ and $\boldsymbol{u}$.

Note the key to the method is the ability to decouple the local and global problem and solve them in a partitioned or two-step procedure. Also note the local problem is solved only once, provided the structure of the unit cell is uniform throughout the continuum body, otherwise the local equations must be solved for all points within the global domain for $\boldsymbol{\chi}$ and $\mathbb{C}^H$.

1.1.1   General Equations and Notation

A brief review of the general governing equations for the development of a solution to a linear elastic boundary value problem are given, along with the associated notation. Following which the weak form equations are then introduced for subsequent use in casting the equations into the finite element method.

Governing Equations

The governing equations for linear elastostatics are given, in indicial notation, by the balance of linear and angular momentum

$$\frac{\partial \sigma_{ij}^{\epsilon}}{\partial x_j} + b_i^{\epsilon} = 0$$

$$\sigma_{ij}^{\epsilon} = \sigma_{ji}^{\epsilon} \ ,$$

$\quad$ (1.1.1)

where σ_{ij}^{ϵ} are the components of the stress tensor, b_i^{ϵ} is the body force per unit volume, and the superscript $(\cdot)^{\epsilon}$ denotes the dependence of the quantity on the local microstructure. The strain-displacement relationship is given as

$$\varepsilon_{ij}^{\epsilon} = \frac{1}{2}\left(\frac{\partial u_i^{\epsilon}}{\partial x_j} + \frac{\partial u_j^{\epsilon}}{\partial x_i}\right) \ ,$$

$\quad$ (1.1.2)

where $\varepsilon_{ij}^{\epsilon}$ are components of the strain tensor and u_i^{ϵ} are components of the displacement field. And lastly, for an elastic stress response the constitutive equations are expressed as

$$\sigma_{ij}^{\epsilon} = \mathbb{C}_{ijkl}^{\epsilon}\varepsilon_{kl}^{\epsilon} \ ,$$

$\quad$ (1.1.3)

where $\mathbb{C}_{ijkl}^{\epsilon}$ are the components of the forth rank material tensor, which relates the components of the stress tensor to components of the strain tensor. As an example for isotropy the components of the material tensor can be expressed as $\mathbb{C}_{ijkl}^{\epsilon} = \mu\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + \lambda\delta_{ij}\delta_{kl}$, where δ_{ij} is the Kronecker delta symbol and μ and λ are the Lame constants. Note the "repeated" indices imply a summation over the variable to which the index is applied.

Weak Form of Governing Equations

Consider a linear elastic three-dimensional body with periodic structure, subjected to arbitrary loading and boundary conditions as our model problem. The statement of the weak form of the boundary value problem is written as:

$$(W) \begin{cases} \text{Find the function } \boldsymbol{u}^\epsilon \in \mathcal{S}^\epsilon \text{ such that for all } \boldsymbol{v} \in \mathcal{V} \\ \int_\Omega \dfrac{\partial v_i}{\partial x_j} \mathbb{C}^\epsilon_{ijkl} \dfrac{\partial u_k^\epsilon}{\partial x_l}\, dV = \int_\Omega v_i b_i^\epsilon\, dV + \int_{\Gamma_t} v_i \bar{t}_i\, dS\ , \end{cases} \tag{1.1.4}$$

where $\bar{t}_i$ are the imposed tractions for the global problem, v_i are arbitrary trial functions, and the spaces $\mathcal{S}^\epsilon$ and $\mathcal{V}$ are given as

$$\mathcal{S}^\epsilon = \left\{ \boldsymbol{u}^\epsilon : \Omega \to \mathbb{R}^n \mid \boldsymbol{u}^\epsilon \in H^1(\Omega) \text{ and } \boldsymbol{u}^\epsilon = \overline{\boldsymbol{u}} \text{ on } \Gamma_u \right\} \tag{1.1.5}$$

and

$$\mathcal{V} = \left\{ \boldsymbol{v} : \Omega \to \mathbb{R}^n \mid \boldsymbol{v} \in H^1(\Omega) \text{ and } \boldsymbol{v} = 0 \text{ on } \Gamma_u \right\}\ . \tag{1.1.6}$$

Again the superscript $(\cdot)^\epsilon$ denotes dependence of the quantity on the local quantities. Note $\Omega \subset \mathbb{R}^n (n = 1, 2\ or\ 3)$ with smooth boundary $\partial\Omega$ represents a bounded reference configuration $\mathcal{B}$ for the continuum body. Also we admit the decomposition of the boundary into two parts: $\Gamma_u \subset \partial\Omega$ where the displacement is prescribed as $\boldsymbol{u} = \overline{\boldsymbol{u}}$ and $\Gamma_t \subset \partial\Omega$ where the traction vector is prescribed, subject to

$$\partial\Omega = \overline{\Gamma_u \cup \Gamma_t} \quad \text{and} \quad \Gamma_u \cap \Gamma_t = \emptyset\ . \tag{1.1.7}$$

1.1.2   Microscopic Weak Form

To develop the microscopic weak form equations a double-scale asymptotic expansion of the solution, $\boldsymbol{u}^\epsilon$, is presented and then substituted into the weak form previously developed. The resulting weak form equations are then used to solve for characteristic quantities, $\boldsymbol{\chi}$, and subsequently form the homogenized material moduli, $\mathbb{C}^H$. The finite element arrays are then developed along with the associated boundary conditions. Lastly, an overview of the numerical implementation is discussed and several numerical studies explored.

Asymptotic Expansion

The key concepts to the asymptotic homogenization theory is a double-scale asymptotic expansion of the solution, $\boldsymbol{u}^\epsilon$, in terms of the local and global level expressed as

$$u_i^\epsilon(\boldsymbol{x}) = \epsilon^0 u_i^0(\boldsymbol{x}, \boldsymbol{y}) + \epsilon^1 u_i^1(\boldsymbol{x}, \boldsymbol{y}) + \epsilon^2 u_i^2(\boldsymbol{x}, \boldsymbol{y}) + \cdots \tag{1.1.8}$$

where $\boldsymbol{x}$ is the coordinate scale on the global level, $\boldsymbol{y}$ is the coordinate scale on the local level, and $\epsilon > 0$ is an asymptotic scale factor ($\boldsymbol{y} = \boldsymbol{x}/\epsilon$). It can be shown that

$$u_i^1 = -\chi_i^{kl}(\boldsymbol{x}, \boldsymbol{y}) \frac{\partial u_k^0}{\partial x_l} + \boldsymbol{\xi}(\boldsymbol{x}) \tag{1.1.9}$$

where $\boldsymbol{\xi}(\boldsymbol{x})$ being a constant of integration and $\boldsymbol{\chi}$ is a characteristic function correspond to various pre-strain load conditions. Another result from the algebraic manipulations when forming (1.1.10) is that $\boldsymbol{u}^0(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{u}^0(\boldsymbol{x})$ (*i.e.* the first term in the asymptotic expansion depends on the global scale only).

1.1.3   Local Weak Form

Substitution of (1.1.8) into (1.1.4) yields, after some algebraic manipulation (see [1] for more details), a weak form for the local problem in terms of a characteristic function χ_p^{kl}.

$$(W) \begin{cases} \text{Find the function } \chi^{kl} \in \mathcal{S}^\epsilon \text{ such that for all } \boldsymbol{v} \in \mathcal{V}^\epsilon \\ \displaystyle\int_{\Omega^\epsilon} \frac{\partial v_i}{\partial y_j} \mathbb{C}_{ijpq} \frac{\partial \chi_p^{kl}}{\partial y_q} \, dV = \int_{\Omega^\epsilon} \frac{\partial v_i}{\partial y_j} \mathbb{C}_{ijkl} \, dV \end{cases} \tag{1.1.10}$$

where Ω^ϵ is the domain of the unit cell, the superscripts kl on the characteristic function $\boldsymbol{\chi}$ correspond to various pre-strain load conditions, which are needed to subsequently determine the homogeneous material moduli expressed as

$$\mathbb{C}_{ijkl}^H(\boldsymbol{x}) = \frac{1}{\Omega^\epsilon} \int_{\Omega^\epsilon} \left(\mathbb{C}_{ijkl} - \mathbb{C}_{ijpq} \frac{\partial \chi_p^{kl}}{\partial y_q} \right) \, dV \ . \tag{1.1.11}$$

1.1.4   Local Stress Computation

Given the states to both the micro and macro level problems the local stress state may be computed. The stress state is obtained, for the local problem, using a first order approximation expressed as

$$\sigma_{ij} = \left[\mathbb{C}_{ijkl} - \mathbb{C}_{ijmn} \frac{\partial \chi_m^{kl}}{\partial y_n} \right] \frac{\partial u_k}{\partial x_l} \ . \tag{1.1.12}$$

1.2   Numerical Developments

The theoretical developments above are recast into a matrix notation to ease implementation into the finite element method. The essential developments include introduction of interpolation functions for the primary variables followed by substitution into the weak form expression.

1.2.1   Finite Element Form for the Local Weak Equations

The finite element arrays are constructed by introducing a set of compatible interpolation functions for the characteristic quantity, $\boldsymbol{\chi}$ and the arbitrary trial functions, $\boldsymbol{v}$ in terms of nodal quantities on the element. These approximate forms are then introduced into the weak form equations to yield a set of approximate equilibrium equations over the element.

Consider the standard isoparametric interpolation functions (see [2] or [3] for more details) for the characteristic quantity, $\boldsymbol{\chi}$ as

$$\chi_i^{kl} = N_I d_i^{I\,(kl)} \qquad (\boldsymbol{\chi}^\alpha = \boldsymbol{N} \boldsymbol{d}^\alpha) \tag{1.2.1}$$

where d_i^I represents the nodal quantity for $\boldsymbol{\chi}$ at the Ith node in the ith coordinate direction and kl (α) corresponds to the loading state imposed (*e.g.* imposed pre-strain in the local 11 direction). Also note for simplicity it is assumed that the approximate interpolation is valid for a particular load condition kl. The interpolation for the arbitrary trial function has the form

$$v_i = N_I \eta_i^I \qquad (\boldsymbol{v} = \boldsymbol{N} \boldsymbol{\eta}) \tag{1.2.2}$$

where η_i^I represents the nodal quantity for $\boldsymbol{v}$ at the Ith node in the ith coordinate direction. The gradients for the functions in equations (1.2.1) and (1.2.2) may be represented as

$$\frac{\partial \chi_i^{kl}}{\partial y_j} = B_I^j d_i^{I\,(kl)} \quad (\nabla \boldsymbol{\chi}^\alpha = \boldsymbol{B} \boldsymbol{d}^\alpha) \quad \text{and} \quad \frac{\partial v_i}{\partial y_j} = B_I^j \eta_i^I \quad (\nabla \boldsymbol{v} = \boldsymbol{B} \boldsymbol{\eta}) \,, \tag{1.2.3}$$

where $\boldsymbol{B}$ is the standard gradient operator, which may be expressed in two-dimensions as

$$\boldsymbol{B} = \begin{bmatrix} N_{1,1} & 0 & \vdots & N_{2,1} & 0 & \vdots & \cdots & \vdots & N_{n,1} & 0 \\ 0 & N_{1,2} & \vdots & 0 & N_{2,2} & \vdots & \cdots & \vdots & 0 & N_{n,2} \\ N_{1,2} & N_{1,1} & \vdots & N_{2,2} & N_{2,1} & \vdots & \cdots & \vdots & N_{n,2} & N_{n,1} \end{bmatrix} \tag{1.2.4}$$

where n is the number of nodes on the element. The material moduli may also be expressed in matrix form as

$$\mathbb{C} = \begin{bmatrix} C_{11} & C_{12} & 0 \\ C_{12} & C_{22} & 0 \\ 0 & 0 & C_{33} \end{bmatrix} \,. \tag{1.2.5}$$

Substituting (1.2.1)-(1.2.3) into (1.1.10) yields an approximate weak form

$$(W^h) \begin{cases} \text{Find the function } d_p \in \mathcal{S}^h \text{ such that for all } \eta_i \in \mathcal{V}^h \\ \displaystyle\int_{\Omega^\epsilon} B_I^j \eta_i^I \mathbb{C}_{ijpq} B_J^q d_p^{J\,(kl)} \, dV = \int_{\Omega^\epsilon} B_I^j \eta_i^I \mathbb{C}_{ijkl} \, dV \,, \end{cases} \tag{1.2.6}$$

where $\mathcal{S}^h \subset \mathcal{S}$ and $\mathcal{V}^h \subset \mathcal{V}$ are the approximate finite element spaces. To further simplify the approximate weak form, W^h may be cast into a direct matrix form as

$$\int_{\Omega^\epsilon} (\boldsymbol{B}\boldsymbol{\eta})^T \, \mathbb{C} \, (\boldsymbol{B}\boldsymbol{d}^\alpha) \ dV = \int_{\Omega\epsilon} (\boldsymbol{B}\boldsymbol{\eta})^T \, \mathbb{C}^\alpha \ dV \ . \tag{1.2.7}$$

where the superscript α on the material moduli matrix denotes the column of the matrix to use for the given loading condition (*e.g.* for loading in the local 11 direction $\alpha = 1$, therefore the first column of $\mathbb{C}$ would be used). Noting that $\boldsymbol{d}$ and $\boldsymbol{\eta}$ are nodal quantities and are independent of the limits of integration and also that $\boldsymbol{\eta}$ is an arbitrary quantity results in

$$\int_{\Omega^\epsilon} \boldsymbol{B}^T \mathbb{C} \boldsymbol{B} \ dV \ \boldsymbol{d}^\alpha = \int_{\Omega\epsilon} \boldsymbol{B}^T \mathbb{C}^\alpha \ dV \ . \tag{1.2.8}$$

Introduce the following notation

$$\boldsymbol{K}_e = \int_{\Omega^\epsilon} \boldsymbol{B}^T \mathbb{C} \boldsymbol{B} \ dV \qquad \text{and} \qquad \boldsymbol{F}_e^\alpha = \int_{\Omega^\epsilon} \boldsymbol{B}^T \mathbb{C}^\alpha \ dV \ , \tag{1.2.9}$$

where $\boldsymbol{K}_e$ and $\boldsymbol{F}_e$ are the element stiffness matrix and load vector. The global stiffness matrix and load vector are determined by assembling the contribution from each element

$$\boldsymbol{K} = \mathbf{A}_e \, \boldsymbol{K}_e \qquad \text{and} \qquad \boldsymbol{F}^\alpha = \mathbf{A}_e \, \boldsymbol{F}_e^\alpha \tag{1.2.10}$$

where $\mathbf{A}$ is the standard assembly operator. After the global arrays are assembled the local periodic boundary conditions are imposed and the reduced system of equations solved for the nodal quantity $\boldsymbol{d}$ as

$$\boldsymbol{K}\boldsymbol{d}^\alpha = \boldsymbol{F}^\alpha \ . \tag{1.2.11}$$

Using the solution computed from (1.2.11) the homogenized material moduli may be computed from (1.1.11) as

$$\mathbb{C}^H = \frac{1}{\Omega^\epsilon} \int_{\Omega^\epsilon} \mathbb{C} - \mathbb{C}\boldsymbol{B}\boldsymbol{d}^* \ dV \ . \tag{1.2.12}$$

Note the nodal solution vector will contain the appropriate solution states corresponding to all possible pre-strain loading conditions. For example in two-dimensions

$$\boldsymbol{d}^* = \begin{bmatrix} \boldsymbol{d}_1^{(1)} & \vdots & \boldsymbol{d}_1^{(2)} & \vdots & \boldsymbol{d}_1^{(3)} \\ \boldsymbol{d}_2^{(1)} & \vdots & \boldsymbol{d}_2^{(2)} & \vdots & \boldsymbol{d}_2^{(3)} \end{bmatrix} \ , \tag{1.2.13}$$

where $\boldsymbol{d}_i^{(1)}$ corresponds to the solution for an imposed pre-strain in the local 11 direction (*i.e.* ε_{11}), $\boldsymbol{d}_i^{(2)}$ corresponds to the solution for an imposed pre-strain in the local 22 direction (*i.e.* ε_{22}), and $\boldsymbol{d}_i^{(3)}$ corresponds to the solution for an imposed pre-strain in the local 12 direction (*i.e.* γ_{12}). Note in three-dimension six pre-strain states must be imposed to enable the determination of all homogenized coefficients.

1.2.2   Periodic Boundary Conditions

The derivation of the required boundary conditions for the solution of (1.2.11) is explained in detail by [1]. The development is introduced and summarized here because it is essential for the solution of the problem. Recall the domain for this local problem is a unit cell of the assumed periodic material. The desired local boundary conditions are simply derived from the periodicity and symmetry of the unit cell.

For simplicity the present discussion is restricted to the two-dimensional case. Recall that the local scale is $\boldsymbol{y}$ and the sides of the unit cell are assumed to be parallel to the $\{y_1, y_2\}$ axes. In addition define the unit cell lengths or periods to be subsequently $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. One can exploit the periodicity and write an expression for the cell displacements, in two-dimensions, as

$$u(y_1, y_2) = u(y_1 + Y_1, y_2) = u(y_1 + Y_1, y_2 + Y_2) = u(y_1, y_2 + Y_2) \tag{1.2.14}$$

$$v(y_1, y_2) = v(y_1 + Y_1, y_2) = v(y_1 + Y_1, y_2 + Y_2) = v(y_1, y_2 + Y_2) \tag{1.2.15}$$

where u and v are the scalar displacements in the y_1 and y_2 directions, respectively. For the discretized domain this implies that pairs of nodes along edges on opposite sides of the unit cell have identical displacements. One approach to implement these conditions would be to write constraint equations that equate the appropriate nodal displacements. Note that care would still need to be taken to eliminate rigid body motion. The more direct approach is to set the boundary conditions by examining the periodicity conditions along with the specific pre-strain load case under consideration. As an example consider the two-dimensional of a unit cell with a circular inclusion. The boundary conditions for the three possible pre-strain load cases are shown in Figure 1.1. Boundary conditions for the three dimensional case would be a simple extension of the two-dimensional case for the remaining normal and two shear pre-strain cases.

1.2.3   Numerical Implementation

To illustrate the effectiveness of the homogenization theory presented the developments were implemented into FEAP, a general purpose finite element program [4]. The elements developed were a linear elastic quadrilateral and a linear elastic hexahedral. For the development of a user element in FEAP the following essential ingredients are needed:

1. Read in material data (call utility routine in FEAP).

2. Construct shape functions (call utility routine in FEAP).

3. Given the current element displacement vector, $\boldsymbol{d}$, construct the stress components, σ_{ij}, and material moduli, $\mathbb{C}_{ijkl}$, at all quadrature points within the element.
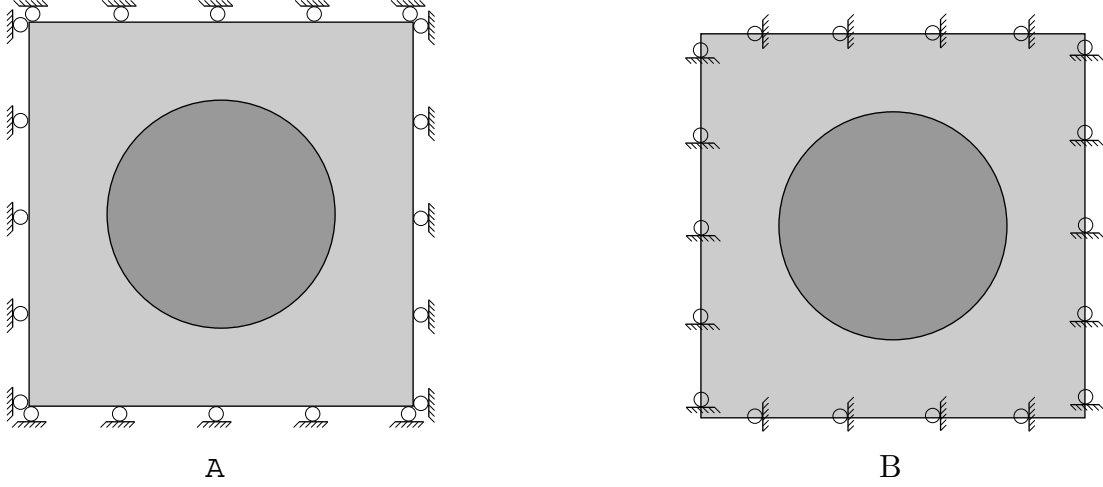
Figure 1.1: Boundary conditions for two-dimensional pre-strain load cases for a unit cell with a circular inclusion. Figure A represents the boundary conditions for the pre-strain load cases: ε_{11} and ε_{22}, while Figure B represents the boundary conditions for the pre-strain load cases: γ_{12}.

4. Given the stress state, $\boldsymbol{\sigma}$, and the material moduli, $\mathbb{C}$, form the residual (*i.e.* internal and external force vectors), given as

$$\boldsymbol{R} = \boldsymbol{f}_{int}(\boldsymbol{d}, \boldsymbol{\alpha}) - \boldsymbol{f}_{ext}(\boldsymbol{d}) \tag{1.2.16}$$

where

$$\boldsymbol{f}_{int} = \int_{\Omega^e} \boldsymbol{B}^T \boldsymbol{\sigma}\ dV \qquad \text{and} \qquad \boldsymbol{f}_{ext}^{\alpha} = \int_{\Omega^e} \boldsymbol{B}^T \mathbb{C}^{\alpha}\ dV\ . \tag{1.2.17}$$

5. Given the material moduli, $\mathbb{C}$, form the tangent (*i.e.* tangent matrix), given as

$$\boldsymbol{K} = \int_{\Omega^e} \boldsymbol{B}^T\ \mathbb{C}\ \boldsymbol{B}\ dV\ . \tag{1.2.18}$$

6. An additional step is the determination of the homogenized moduli after the solution vector, $\boldsymbol{d}^{\alpha}$, has been determined, given as

$$\mathbb{C}^H = \frac{1}{\Omega^\epsilon} \int_{\Omega^\epsilon} \mathbb{C} - \mathbb{C}\boldsymbol{B}\boldsymbol{d}^*\ dV\ , \tag{1.2.19}$$

where $\boldsymbol{d}^*$ is given in (1.2.13).

The only arrays, which are output from the user element are the element residual, $\boldsymbol{R}$, and tangent, $\boldsymbol{K}$. FEAP assembles the individual element contributions and forms the global

arrays. Since FEAP is inherently a non-linear code, the solution is determined using a Newton-Raphson procedure. Note since the problem under consideration is linear only one iteration is needed to obtain the solution. A listing of the elements is located in Appendix A. Several simulations are presented in subsequent sections to show the robustness of the formulation in comparison with analytical results.

1.2.4   Global Weak Form

The resulting homogenized material moduli is then cast into a macroscopic weak form, which will enable the solution of the global variables, $\boldsymbol{u}$.

Given the homogenized material moduli the weak form for the global problem may be cast as

$$(W) \begin{cases} \text{Find the function } u_k \equiv u_k^0 \in \mathcal{S} \text{ such that for all } \boldsymbol{v} \in \mathcal{V} \\ \displaystyle\int_\Omega \frac{\partial v_i}{\partial x_j} \mathbb{C}_{ijkl}^H \frac{\partial u_k}{\partial x_l} \, dV = \int_\Omega v_i f_i \, dV + \int_{\Gamma_t} v_i \bar{t}_i \, dS \ , \end{cases} \tag{1.2.20}$$

where f_i is the average body force per unit volume given as

$$f_i(\boldsymbol{x}) = \frac{1}{\Omega^\epsilon} \int_{\Omega^\epsilon} b_i \, dV \ . \tag{1.2.21}$$

1.2.5   Finite Element Form for the Global Weak Equations

As before the finite element arrays are constructed by introducing a set of compatible interpolation functions for the displacement field, $\boldsymbol{u}$ and the arbitrary trial functions, $\boldsymbol{v}$ in terms of nodal quantities on the element. These approximate forms are then introduced into the weak form equations to yield a set of approximate equilibrium equations over the element.

Consider the following interpolation functions

$$u_i = N_I d_i^I \quad (\boldsymbol{u} = \boldsymbol{N}\boldsymbol{d}) \qquad \text{and} \qquad v_i = N_I \eta_i^I \quad (\boldsymbol{v} = \boldsymbol{N}\boldsymbol{\eta}) \ , \tag{1.2.22}$$

where the syntax follows from the previous section. The gradients for (1.2.22) are expressed as

$$\frac{\partial u_i}{\partial y_j} = B_I^j d_i^I \quad (\nabla\boldsymbol{u} = \boldsymbol{B}\boldsymbol{d}) \qquad \text{and} \qquad \frac{\partial v_i}{\partial y_j} = B_I^j \eta_i^I \quad (\nabla\boldsymbol{v} = \boldsymbol{B}\boldsymbol{\eta}) \ . \tag{1.2.23}$$

Casting (1.2.22)-(1.2.23) into (1.2.20) yields an equivalent matrix form for the approximate weak form equations, again noting that $\boldsymbol{d}$ and $\boldsymbol{\eta}$ are nodal quantities and are independent

of the limits of integration and also that $\boldsymbol{\eta}$ is an arbitrary quantity

$$\underbrace{\int_{\Omega} \boldsymbol{B}^T \mathbb{C}^H \boldsymbol{B} \ dV}_{\boldsymbol{K}_e} \ \boldsymbol{d} = \underbrace{\int_{\Omega} \boldsymbol{N}^T \boldsymbol{f} \ dV + \int_{\Gamma_t} \boldsymbol{N}^T \overline{\boldsymbol{t}} \ dS}_{\boldsymbol{F}_e} \ . \tag{1.2.24}$$

The resulting element arrays, $\boldsymbol{K}_e$ and $\boldsymbol{F}_e$ are assembled to form global arrays from which global boundary conditions are imposed and the nodal displacement vector, $\boldsymbol{d}$ determined.

1.2.6   Overview

To summarize the previous sections the following steps are implemented:

1. Develop a mathematical model for the local periodic unit cell

   (a) Discretize the domain of the unit cell

   (b) Form the element arrays $\boldsymbol{K}_e$ and $\boldsymbol{F}_e^\alpha$, see equation (1.2.9)

   (c) Assemble the element arrays into global arrays, see equation (1.2.10)

   (d) Impose periodic boundary conditions and pre-strain conditions α

   (e) Solve the reduced system of algebraic equations for the nodal quantity $\boldsymbol{d}$, see equation (1.2.11)

   (f) After solving for all possible pre-strain conditions, α, form the homogenized material moduli, $\mathbb{C}^H$, for the periodic unit cell, see equation (1.2.12)

2. Develop a mathematical model for the global problem

   (a) Discretize the physical global domain

   (b) Form the element arrays $\boldsymbol{K}_e$ and $\boldsymbol{F}_e$, see equation (1.2.24)

   (c) Assemble the element arrays into global arrays

   (d) Impose boundary conditions and loading conditions

   (e) Solve the reduced system of algebraic equations for the nodal quantity $\boldsymbol{d}$

There are several noteworthy observations obtained from the implementation of the homogenization theory presented in the previous sections:

- Given complex domains for a unit cell the finite element method affords a simple and attractive means for determining the homogenized material moduli for subsequent use in computing the response of a global problem.

- From the convergence study in the following section of a unit cell with a circular fiber, which has a continuous geometry, converges rapidly. It should be noted that a unit cell with the rectangular void, which has high solution gradients at the re-entrant corner, will converge slower as excepted.

- An improvement to the given model (*i.e.* high accuracy for course mesh densities) would be to implement the formulation into the setting of an enhanced finite element formulation, see for example [5] or [6].

- The develop of the homogenized material moduli in three dimensions follows by a simple extension of the finite element model presented.

- A three dimensional representation of the homogenization element will yield the input needed for either a continuum based element or a multi-layered composite shell element needed to solve composite systems with material and geometric periodicity accurately and efficiently.

Section 2

Numerical Simulations

2.1 Convergence Study

To aid in determining the element type and mesh density for computing the homogenized material moduli for subsequent unit cells several representative simulations were performed in two-dimensions outlined below.

2.1.1 Model Configurations

The cell configuration under consideration was a circular fiber surrounded by a unit square matrix, see Figure 2.1. For the present study the several variations of the configuration (*i.e.* fiber diameter) were explored. Table 2.1 list the associated parameters for each simulation subsequently performed.

Table 2.1: Properties for the Unit Cell Matrix with an Interior Circular Fiber.

Quantity	Matrix	Fiber
Elastic Modulus (GPa)	$E_m = 4.62$	$E_f = 23.1$
Poisson's ratio	$\nu_m = 0.36$	$\nu_f = 0.4$

2.1.2 Simulations

To determine the most efficient element formulation and mesh density for the determination of the material moduli a convergence study was performed for the various configuration
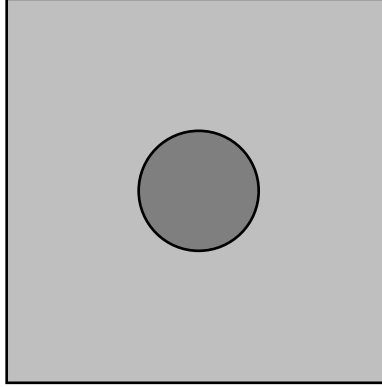
Figure 2.1: Two-dimensional representation of the unit cell configurations examined. Specifically, a unit cell with a circular fiber embedded in a matrix.

investigated. The results from the convergence study will enable the determination of the uncertainty for the results for a given element formulation and mesh density. Note if the global problem is not spatially homogeneous then a separate analysis will need to be performed at each point (*i.e.* Gauss point) within the global mesh discretization and thus efficiency and robustness is key to making the homogenization technique feasible on large scale linear problems or with the extension to the non-linear regime. The element formulations used for the convergence study are:

Q4 : Four noded quadrilateral using a standard displacement based formulation and a standard 4-pt Gaussian quadrature rule, see [2] or [3].

Q8 : Eight noded quadrilateral using a standard displacement based formulation and a standard 9-pt Gaussian quadrature rule, see [2] or [3].

Q9 : Nine noded quadrilateral using a standard displacement based formulation and a standard 9-pt Gaussian quadrature rule, see [2] or [3].

ME2 : Four noded mixed-enhanced quadrilateral using a standard 4-pt Gaussian quadrature rule, see [6].

The mesh densities considered are structured meshes with several controlling parameters which affect the density. Figure 2.2 depicts several typical mesh discretizations. Utilizing the element formulations given in Chapter 1 a convergence study of the homogenized material moduli coefficients was performed. Figure 2.3 depicts a convergence study of a unit cell with an interior circular fiber with a radius $r = 0.3$ (fiber volume ratio $V_f = 0.283$), while Figure 2.4 depicts a convergence study of a unit cell with an interior circular fiber with a radius $r = 0.4$ (fiber volume ratio $V_f = 0.503$).
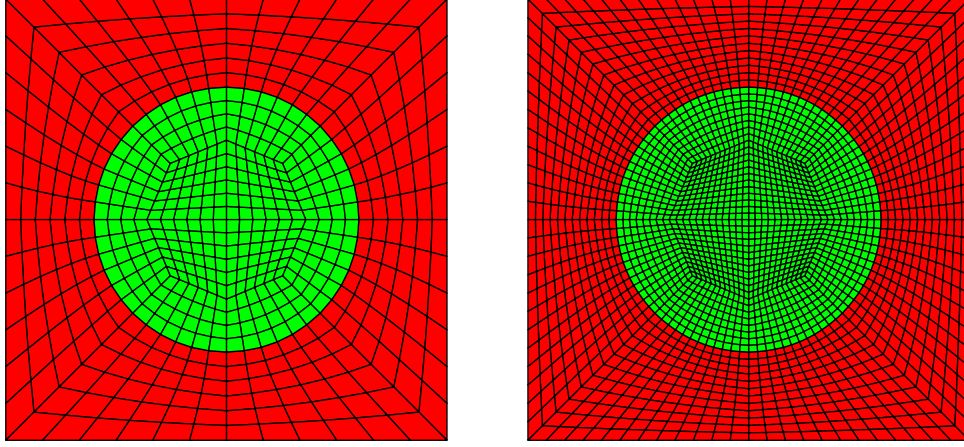
Figure 2.2: Typical mesh discretizations for the unit cell with a circular fiber.

Remark 2.1.

1. For the convergence studies performed the range of the ordinate on the figures was limited to $D_{ij}^{avg}(1 \pm 0.1)$, where D_{ij}^{avg} is the average value of the coefficient D_{ij}. Thus for the studies presented all simulations were within approximately 5 parts in 1000 of the numerically convergent solution. Hence a very course mesh may be utilized to obtained the homogenized material moduli coefficients, since for most engineering application the material moduli are only known to within 50 parts in 1000, at best.

2. Utilizing the course mesh approximation considerable computation saving can be obtained. For applications where spatially non-homogeneous properties or non-linear regimes exist great saving can be obtained and large scale problem then become realizable.
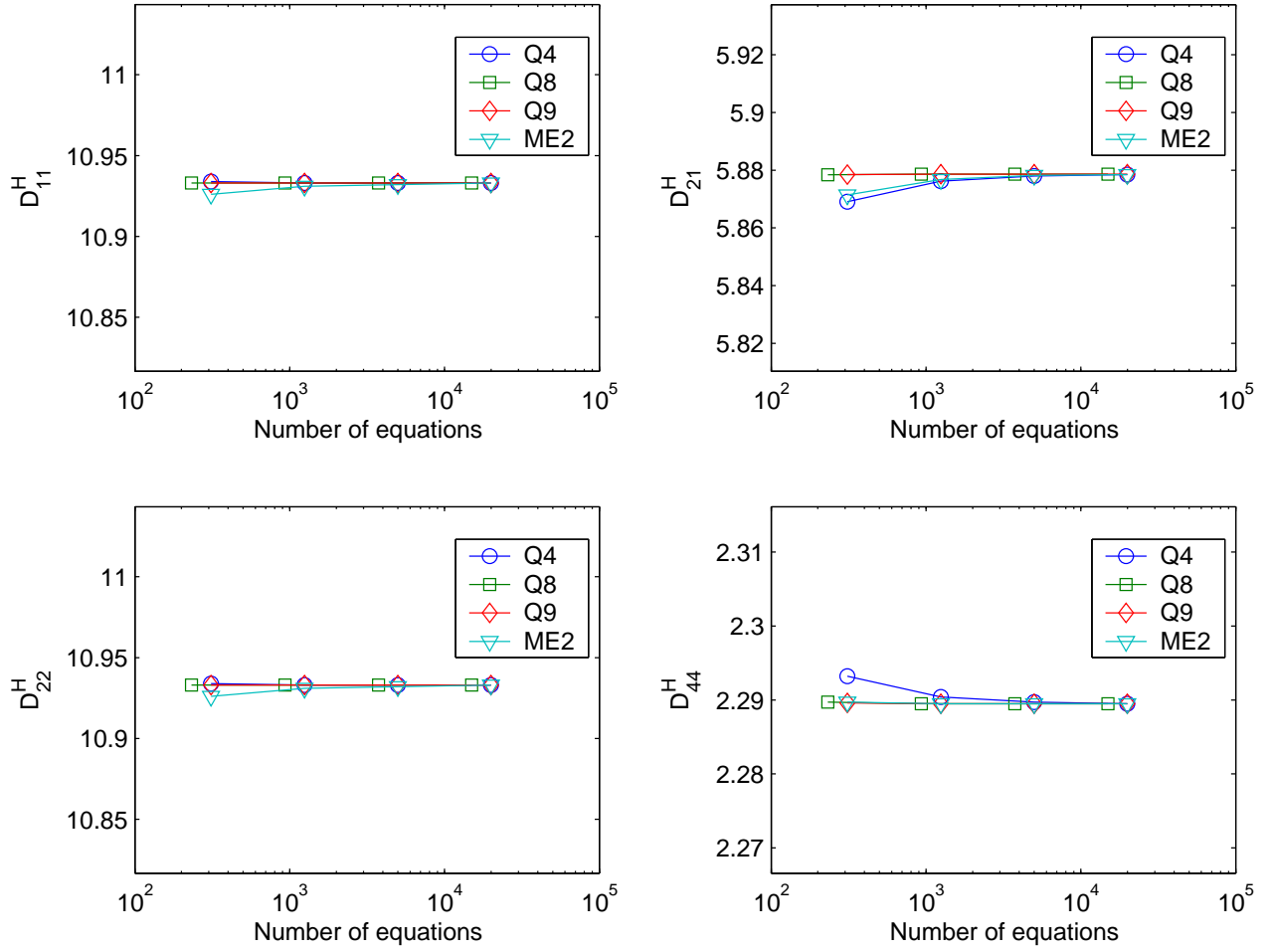
Figure 2.3: Convergence study of a unit cell with a circular fiber with a radius of $r = 0.3$.
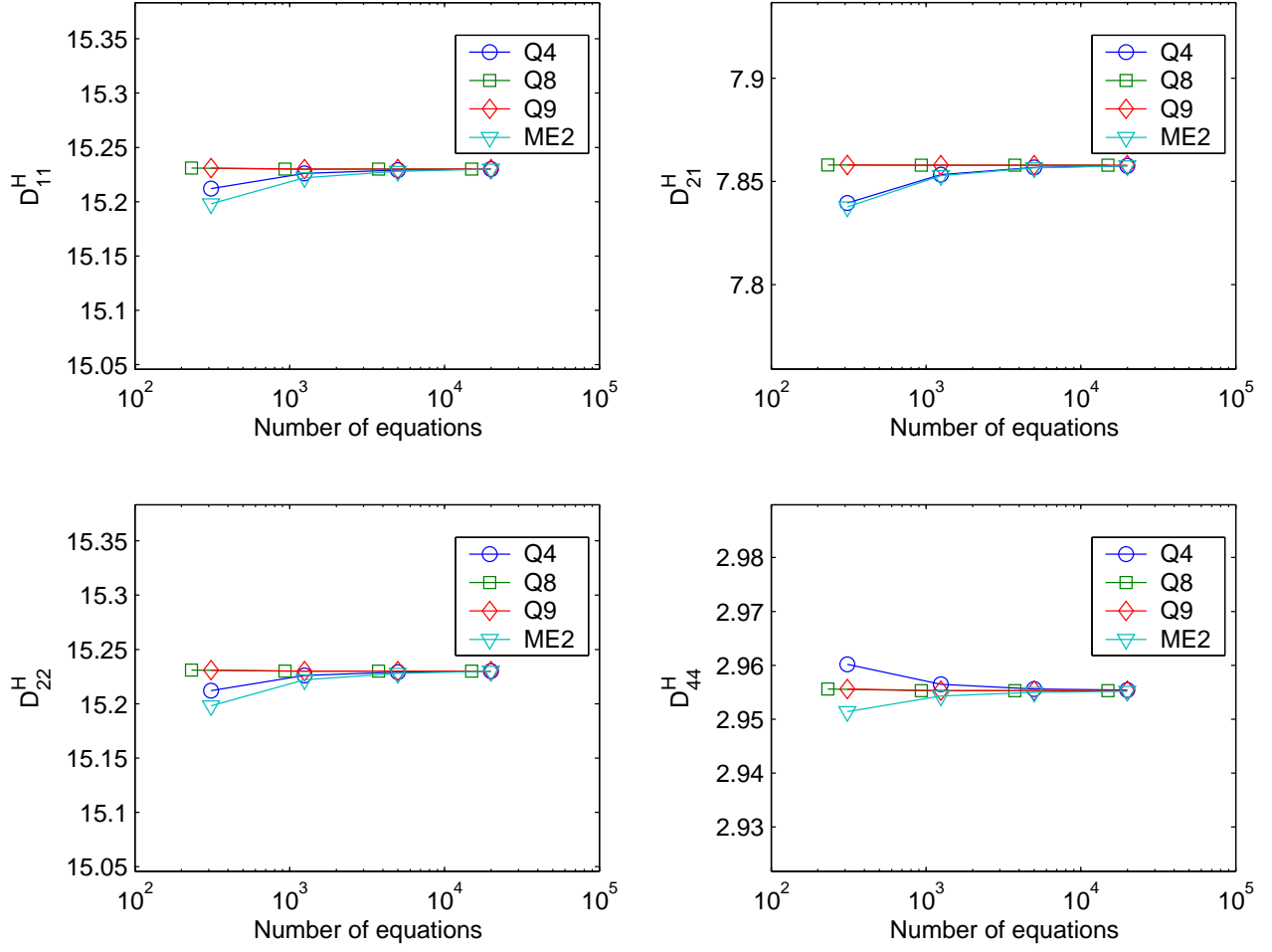
Figure 2.4: Convergence study of a unit cell with a circular fiber with a radius of $r = 0.4$.

2.2   Two-Dimensional Simulations

To aid in comparing the robustness of the homogenization method with more classical approaches, such as micromechanics [7] and [8] several representative unit cell configurations in both two and three dimensions were considered. In this section the two dimensional results are explored, while the following section gives results for three dimensions. For the various configurations the resulting material moduli are tabulated.

2.2.1   Model Configurations

The model cell configuration examined is a circular fiber surrounded by a unit square matrix, see Figure 2.5. For the present study several variations of the configuration (*i.e.* fiber volume ratio and material constituents) were explored. Tables 2.2 and 2.3 list the associated parameters for each simulation performed.
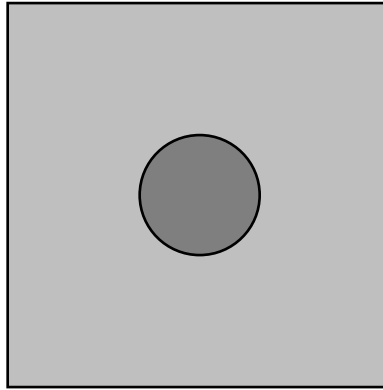
Figure 2.5: Two-dimensional representation of the unit cell configurations examined. Specifically, a unit cell with a circular fiber embedded in a matrix.

Table 2.2: Simulation #1: Properties for the unit cell matrix with an interior circular fiber. The fiber is E-Glass 21 x K43 Gevetex and the matrix is LY556/HT907/DY063 Epoxy. The volume fraction or fiber volume ratio is $V_f = 0.62$. Note both the fiber and the matrix are considered isotropic.

Quantity	Matrix: Epoxy	Fiber: E-Glass
Elastic Modulus (GPa)	$E = 3.35$	$E = 80.00$
Poisson's ratio	$\nu = 0.35$	$\nu = 0.20$

Table 2.3: Simulation #2: Properties for the unit cell matrix with an interior circular fiber. The fiber is AS4 and the matrix is T700/RS1 Epoxy. The volume fraction or fiber volume ratio is $V_f = 0.52$. Note the matrix is considered isotropic, while the fiber is transversely isotropic. In addition the "1" direction is the fiber direction and the "2" and "3" directions are the transverse to the fiber direction.

Quantity	Matrix: Epoxy	Fiber: AS4
Elastic Modulus (msi)	$E = 3.35$	$E_{11} = 33.80$
		$E_{22} = 2.19$
		$G_{12} = 1.30$
		$G_{23} = 4.90$
Poisson's ratio	$\nu = 0.35$	$\nu_{12} = 0.28$
		$\nu_{23} = 0.49$

2.2.2   Moduli Results

For the subsequent developments the typical discretization of a circular fiber within a matrix is shown in Figure 2.6 was utilized for the finite element method implementation of the homogenization theory. The analysis of the unit cell was performed for both plane strain and plane stress. Throughout the kinetic and kinematics are assumed linear. To aid in validating the homogenization results the components of the material stiffness matrix $\mathbf{D}$ are used to compare the numerical and analytical results, where the ordering for the three-dimensional state is given as

$$
\begin{Bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{23} \\ \sigma_{13} \\ \sigma_{12} \end{Bmatrix} = \begin{bmatrix} D_{11} & D_{12} & D_{13} & 0 & 0 & 0 \\ D_{21} & D_{22} & D_{23} & 0 & 0 & 0 \\ D_{31} & D_{32} & D_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & D_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & D_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & D_{66} \end{bmatrix} \begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ \varepsilon_{23} \\ \varepsilon_{13} \\ \varepsilon_{12} \end{Bmatrix} \tag{2.2.1}
$$

where the "1" direction denotes the fiber direction and the "2" and "3" directions denote transverse to the fiber direction. The results of the finite element homogenization results and micromechanics for the stiffness matrix are shown in Tables 2.4-2.5.

Figure 2.6: Quarter symmetry of a unit cell with a circular fiber embedded in a matrix.

Table 2.4: Analysis results for the unit cell matrix with an interior circular fiber, Simulation #1: E-Glass/Epoxy.

Analysis	D_{22}	D_{33}	D_{23}	D_{32}	D_{44}
Homogenization					
Plane Stress	13.743	13.743	2.492	2.492	3.437
Plane Strain	17.936	17.936	4.971	4.971	3.467
Micromechanics	15.068	15.068	6.944	6.944	4.061

Table 2.5: Analysis results for the unit cell matrix with an interior circular fiber, Simulation #2: AS4/Epoxy.

Analysis	D_{22}	D_{33}	D_{23}	D_{32}	D_{44}
Homogenization					
Plane Stress	1.212	1.212	0.425	0.425	0.326
Plane Strain	2.077	2.077	1.252	1.252	0.329
Micromechanics	1.243	1.243	0.631	0.631	0.310

Remark 2.2.

- From Tables 2.4 and 2.5 we see that the D_{22} and D_{33} moduli terms bound the three dimensional micromechanics solution, as expected. Whereas, the remaining terms do not yield accurate results due to the boundary value problem used to developed the analytical micromechanics solution.

2.3   Three-Dimensional Simulations

In this section the models presented in the previous section are revisited in a three dimensional setting. As before for the various configurations the resulting material moduli are tabulated.

2.3.1   Model Configurations

The model cell configuration examined is a circular fiber surrounded by a unit square matrix and a hexagonal pack matrix, see Figure 2.7. For the present study several variations of the configuration (*i.e.* fiber volume ratio and material constituents) were explored. Tables 2.2 and 2.3 list the associated parameters for each simulation performed.
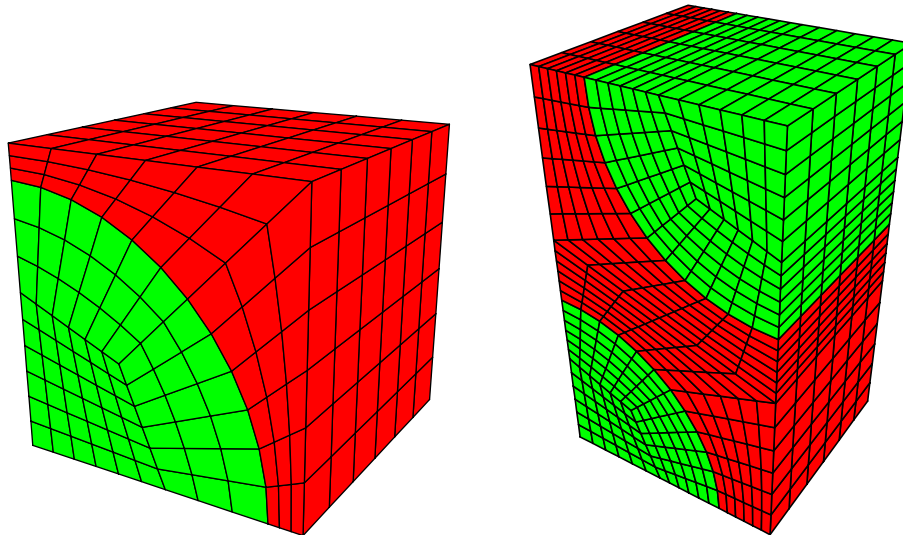
Figure 2.7: Three-dimensional representation of the unit cell configurations examined. Specifically, a unit cell with a circular fiber embedded in a square and hexagonal matrix.

2.3.2   Moduli Results

For the subsequent developments the typical discretization of a circular fiber within a matrix is shown in Figure 2.7 was utilized for the finite element method implementation of the homogenization theory. Throughout the kinetic and kinematics are assumed linear. The material parameters from the numerical homogenization and analytical methods are compared to aid in validating the proposed methodology. The results of the finite element homogenization results and micromechanics for the stiffness matrix are shown in Tables **??-??** below.
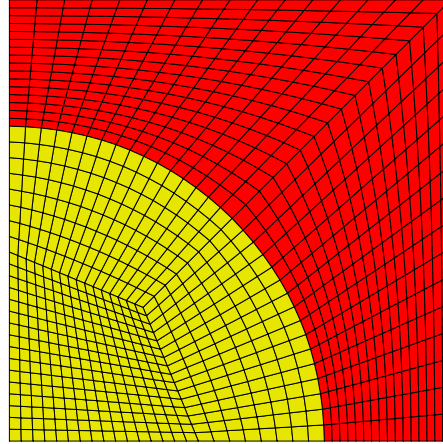
Table 2.6: Analysis results for the unit cell matrix with an interior circular fiber, Simulation #1: E-Glass/Epoxy.

Analysis	E_{11}	E_{22}	G_{12}	G_{23}	ν_{12}
Soden	53.48	17.7	5.83	na	0.278
Hashin	50.9	11.699	4.602	4.061	0.249
Christensen	50.9	12.869	4.605	4.647	0.249
Homogenization					
(square pack)	50.77	16.317	4.922	3.490	0.247
(hex pack)	50.73	12.607	4.516	4.513	0.249

Table 2.7: Analysis results for the unit cell matrix with an interior circular fiber, Simulation #2: T700/RS 1 Epoxy.

Analysis	E_{11}	E_{22}	G_{12}	G_{23}	ν_{12}
Hashin	17.785	0.9177	0.38007	0.30581	0.3097
Homogenization					
(square pack)	17.735	1.0135	0.3837	0.3545	0.3095
(hex pack)	17.724	0.933	0.372	0.436	0.3098

Remark 2.3.

- The difference between the experimental E_{11} and all analytical methods illustrates the problem of fiber constituent data

- The hex pack homogenization results correlate well with accepted analytical models

- Square pack results are suspect in some instances which is consistent with hex pack preference of many authors

Section 3

Conclusions

There are several noteworthy observations obtained from the implementation of the homogenization theory presented in the previous sections:

- Given complex domains for a unit cell the finite element method with a numerical implementation of the homogenization theory affords a simple and attractive means for determining the homogenized material moduli for subsequent use in computing the response of a global problem.

- From the convergence study performed for the unit cell with a circular fiber with continuous geometry, the moduli results converges rapidly (*i.e.* high accuracy for course mesh densities).

- The development of the homogenized material moduli in three dimensions follows by a simple extension of the two dimensional finite element model presented in Section 1

- A three dimensional representation of the homogenization element will yield the input needed for the multi-layered composite shell structures that are used in both military and non-military applications. Which could be used to easily design new composite material systems.

- Homogenization is attractive as it derives the elastic constants in a completely general manner. This seems to be an improvement over a series of finite element models of the unit cell and some type of stress or strain averaging.

- There is a need for accurate constituent material data to correctly compare analytical and numerical methods of material property prediction.

Bibliography

[1] B. Hassani and E. Hinton. *Homogenization and Structural Topology Optimization*. Springer, New York, 1999.

[2] T.J.R. Hughes. *The Finite Element Method*. Prentice-Hall, New Jersey, 1987.

[3] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method*, volume 1. Butterworth-Heinemann, Oxford, fifth edition, 2000.

[4] R.L. Taylor. *FEAP - Finite Element Analysis Program*. University of California: Berkeley, http://www.ce.berkeley.edu/∼rlt, 2002.

[5] J.C. Simo and M.S. Rifai. A class of mixed assumed strain methods and the method of incompatible modes. *International Journal for Numerical Methods in Engineering*, 29:1595–1638, 1990.

[6] E.P. Kasper and R.L. Taylor. A mixed-enhanced strain method part I: Geometrically linear problems. *Computers and Structures*, 75(3):237–250, 2000.

[7] R. M. Christensen. *Mechanics of Composite Materials*. John Wiley & Sons, New York, first edition, 1979.

[8] Z. Hashin. Analysis of properites of fibers with anisotropic constituents. *Journal of Applied Mechanics*, pages 543–550, 1979.

Appendices Available Upon Request

Water Treatment with Manganese Oxide Biofilms

Yarrow M. Nelson
Patricia Coffey
Civil and Environmental Engineering

Bunkim Chokshi
Biochemical Engineering

Cal Poly State University

Water Treatment with Manganese Oxide Biofilms
California Central Coast Research Partnership

FINAL REPORT
November 30, 2003

Yarrow M. Nelson, Assoc. Professor, Civil and Environmental Engineering
Bunkim Chokshi, Graduate Student, Biochemical Engineering
Patricia Coffey, Graduate Student, Civil and Environmental Engineering
Cal Poly State University, San Luis Obispo, California

Introduction:

The overall goal of this project is to develop fixed-film bioreactors to test novel water treatment scenarios using biogenic manganese oxide biofilms. Earlier research by the PI showed that biogenic manganese oxides exhibit extremely high specific surface area and reactivity, making them excellent candidates for water treatment applications (Nelson and Lion, 2003). It is likely that these biofilms will be useful for removing toxic metals from contaminated water and also oxidizing recalcitrant organic compounds such as polyaromatic hydrocarbons and chlorinated solvents. Possible applications include industrial wastewater pretreatment to remove toxic metals from electroplating and other industries, bioremediation of contaminated groundwater containing toxic metals, and treatment and remediation of toxic hydrocarbons. To test the feasibility of using manganese oxide biofilms for these applications, we proposed to construct fixed-film bioreactors and inoculate the biofilm reactors with manganese-oxidizing bacteria. These bioreactors are then to be used for testing toxic metal adsorption by the biofilms and their ability to oxidized recalcitrant compounds such as polyaromatic hydrocarbons and chlorinated solvents.

This research project was selected as an alternate C3RP proposal, so funding through ONR began in May 2003 with $6,000 out of the requested $27,000 budget. This report describes the research accomplishments made in this short time frame, which lay the foundation for completion of the research next year. These accomplishments included construction of the biofilm reactors along with all peripheral equipment as well as procurement and culturing of the manganese-oxidizing bacteria. Ultimately, the biofilm systems will be tested for treatment of toxic metals and organic compounds in industrial wastewater. Preliminary accomplishments for this period are described below.

Background:

Biologically oxidized manganese is produced when certain species of bacteria enzymatically oxidize manganese from the soluble reduced form (Mn^{2+}) to the solid-phase manganese oxide (MnO_2). Manganese oxide minerals have been shown to be efficient at adsorbing toxic metals out of solution, making them a viable candidate for removing toxic metals from polluted water. Almost all previous work in this area has been conducted using manganese oxide minerals that were formed abiotically (chemical reaction as opposed to biological reaction). However, the PI recently discovered that

1

manganese oxides prepared by biological oxidation exhibit an order of magnitude greater surface area than the abiotic Mn oxide, and therefore their adsorption capacity for toxic metals is an order of magnitude greater than that of the abiotic materials (see Figure 1 and Nelson, et al., 1999). Thus, biogenic manganese oxide is ripe for exploitation for use in a variety of water treatment and remediation applications.

Further, manganese oxides were recently shown to be capable of oxidizing humic acids, which are natural compounds found in water that are very difficult to biodegrade. Since manganese oxides are capable of oxidizing humic acids, they might also be expected to be capable of oxidizing toxic organic compounds such as hydrocarbons and polyaromatic hydrocarbons. The reactivity of biogenic manganese oxides is expected to be even greater, and therefore they are likely to be successful materials for bioremediation of groundwater containing these materials.

Toxic metal pollution continues to be a serious threat to receiving waters in the United States (Nriagu and Pacyna 1988). Toxic metals, such as lead (Pb), cadmium (Cd), chromium (Cr), mercury (Hg), etc., are by-products of many industrial processes, and metal pollution from such facilities often is discharged to municipal wastewater treatment facilities. Since metals are not biodegradable, they typically end up in the sludge at these facilities, making sludge disposal on land problematic. Therefore industrial pretreatment systems for toxic metals need to be developed. The proposed research would develop a very promising, yet currently untested method for removing toxic metals from industrial wastewater by making use of biologically oxidized manganese which has been shown by the author to have a high adsorption capacity for toxic metals (Nelson, Lion et al. 1999). The method of production of these manganese oxides and their trace metal adsorption properties are described below.

Similarly, groundwater contamination by numerous polyaromatic hydrocarbons and chlorinated hydrocarbons is a huge problem in the United States and has led to the designation of thousands of contaminated sites (USEPA 1984), and cleaning up these sites is a multi-billion dollar business (U.S.EPA 1993). There is some preliminary evidence that manganese oxides may be capable of oxidizing these hydrocarbon compounds (see below), but this has not been capitalized on by any research groups as of this writing.
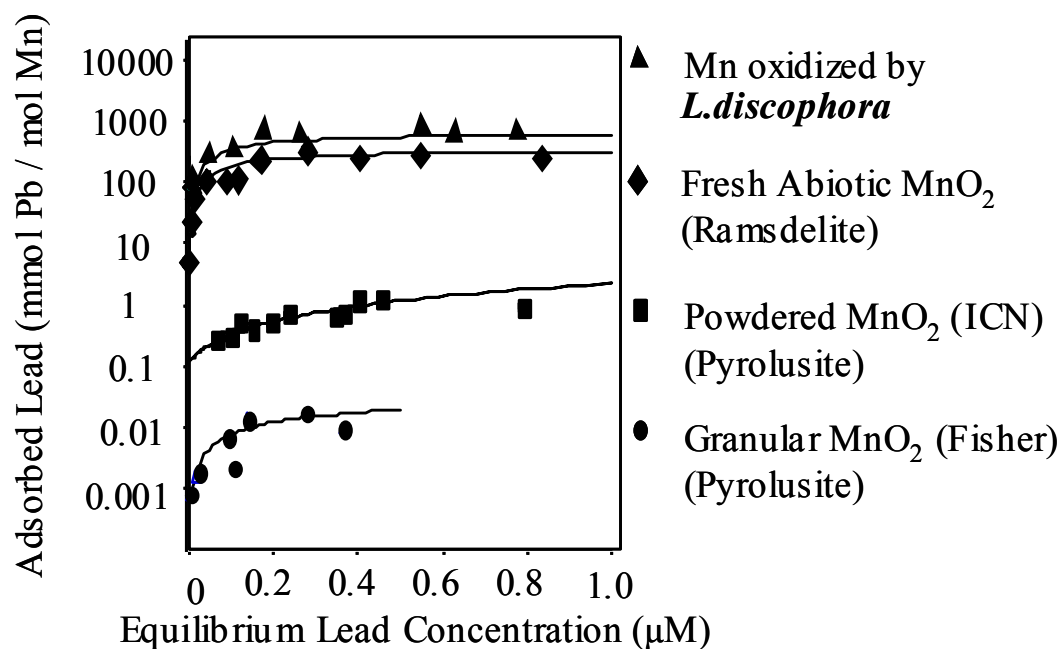
2

Figure 1. Comparison of the lead adsorption capacity of biogenic manganese oxide to the lead adsorption capacity of abiotic manganese oxide minerals. Note that the y-axis is a log scale, and therefore the biogenic manganese oxide adsorbs orders of magnitude more lead at a given concentration than the abiotic minerals.

Manganese oxides are naturally present in soils and sediments in a variety of mineral forms (Bartlett, 1988). These manganese oxides exhibit a wide range of adsorption capacities that are similar or lower than that of iron oxides (Catts and Langmuir 1986). However, the earlier studies on toxic metal adsorption by manganese oxides were conducted exclusively with manganese oxide minerals that were not produced by biological oxidation (abiotic manganese oxides). A few years ago the author prepared biogenic manganese oxides under highly controlled laboratory conditions and measured their toxic metal adsorption capacity. The biogenic manganese oxides were prepared using pure cultures of the Mn-oxidizing bacterium *Leptothrix discophora*. These biogenic Mn oxides exhibited specific surface areas and toxic metal adsorption capacities that were significantly greater than that of abiotic Mn oxides (Figure 1 and Nelson, et al. 1999). The author has subsequently measured biological Mn oxidation kinetics (Zhang et al. 2002) and examined toxic metal adsorption to mixtures of Mn and Fe oxides (Nelson et al. 2002). However, what remains to be done is to prove the application of this technology for the treatment of metal-contaminated wastewaters. Therefore, I propose to produce biofilms containing biogenic Mn oxides in the laboratory and test the ability of these biofilms to remove toxic metals from solution.

There is also some reason to believe that biogenic Mn oxides may be very effective in oxidizing toxic organic compounds. For many years microbiologist were puzzled as to why Mn-oxidizing bacteria oxidize Mn when they appear to obtain no energy from the process (Adams and Ghiorse 1986). The mystery may have been solved when Sunda and Kieber discovered that manganese oxides were capable of oxidizing humic acids (Sunda and Kieber 1994). Humic acids are very recalcitrant natural compounds found in water, and they are generally not digestible by bacteria. Thus, the Mn-oxidizing bacteria may have benefited from their production of Mn oxide by the oxidation of the humic acids into compounds that are bioavailable to the bacteria. The question is, if biogenic Mn oxides can oxidize recalcitrant humic acids, might they also be capable of oxidizing toxic groundwater contaminants. This would be a significant breakthrough in the field of bioremediation, yet to the best of my knowledge, no-one has attempted to capitalize on this capability or even to document it. I therefore propose to test the ability of biogenic Mn oxides to oxidize simple hydrocarbons as model groundwater contaminant compounds.

Preliminary Research Accomplishments:

The overall experimental plan was to grow biofilms of the manganese-oxidizing bacterium *Leptothrix discophora* SS-1 in biofilm reactors, and then introduce dissolved manganese to prepare biofilms containing biogenic Mn oxide. These biofilms are then to be grown in a laboratory bench-scale bioreactor that can be tested for its ability to remove toxic metals from solution and its ability to oxidize toxic organic compounds.

Despite the late start and limited budget, significant progress was made in setting up bioreactor systems for producing the Mn oxide biofilms. Two graduate students (listed above) worked on this project in the summer of 2003, but their funding ran out by Fall Quarter 2003. During this time, the physical bioreactors were constructed, and pure

4

cultures of the Mn-oxidizing bacteria were obtained and grown in our laboratory. Specific accomplishments on this preliminary project are listed below:

Bioreactor construction: Two fixed-film trickling filter bioreactors were designed and constructed in the laboratory (Figure 1). One bioreactor will be used for testing lead adsorption and oxidation of organic compounds, while the second will serve as a manganese-free control. The biofilm reactors are designed to operate with a pure culture of *L. discophora*, and therefore the design incorporates elements to maintain sterile conditions. Each biofilm reactor consists of a 28.5 cm tall by 6.5 cm diameter Plexiglas column packed with 0.5 cm beads (Figure 2). Feed is pumped over the bed through a sprinkler located at the top center of the column which will spray fluid over the entire diameter of the column using a peristaltic pump. Filtered air is sparged (counter-current) into the column using at 2000 cm^3/min and 4.0 psi. Each biofilm reactor is equipped with a 2-L jacketed glass feed reservoir complete with stirring, pH control, temperature control, and dissolved oxygen control.

Pure culture growth of *Leptotrhix Discophora* SS-1
        L. Discophora SS-1 was obtained from American Type Cell Culture (ATCC) for use in the biofilm reactor. The original sample obtained from ATCC did not appear to be viable, so a second sample was obtained from ATCC. This sample also did not appear to be viable, so the PI obtained a fresh culture from researchers at Cornell University who were the original depositors of the ATCC culture. This culture was successfully grown on plates on MSVP medium (mineral salts vitamins pyruvate). Growth on plates containing $MnSO_4$ showed brown colonies, indicative of manganese oxidation.

Graphite furnace atomic absorption spectroscopy
        Once the biofilms are established in the bioreactors, the first test will be of their lead adsorption capacity. Lead concentrations in the simulated wastewater will be measured in the influent and effluent of the biofilm reactor as well as the control. Lead will be measured using graphite furnace atomic absorption spectrophotometry (GFAAS). The graduate students have learned how to use the Perkin Elmer GFAAS and have prepared preliminary calibration curves to test the accuracy and repeatability of lead measurements.

Conclusions:

Construction of two biofilm reactors has been completed, along with all peripheral equipment needed to operate these bioreactors. The difficult task of obtaining and growing pure cultures of a manganese-oxidizing bacterium has also been completed. These major accomplishments lay the foundation for the experiments proposed for next year in which the biofilm reactors will hopefully be tested for their ability to treat water contaminated with toxic metals and recalcitrant organic compounds such as chlorinated solvents.
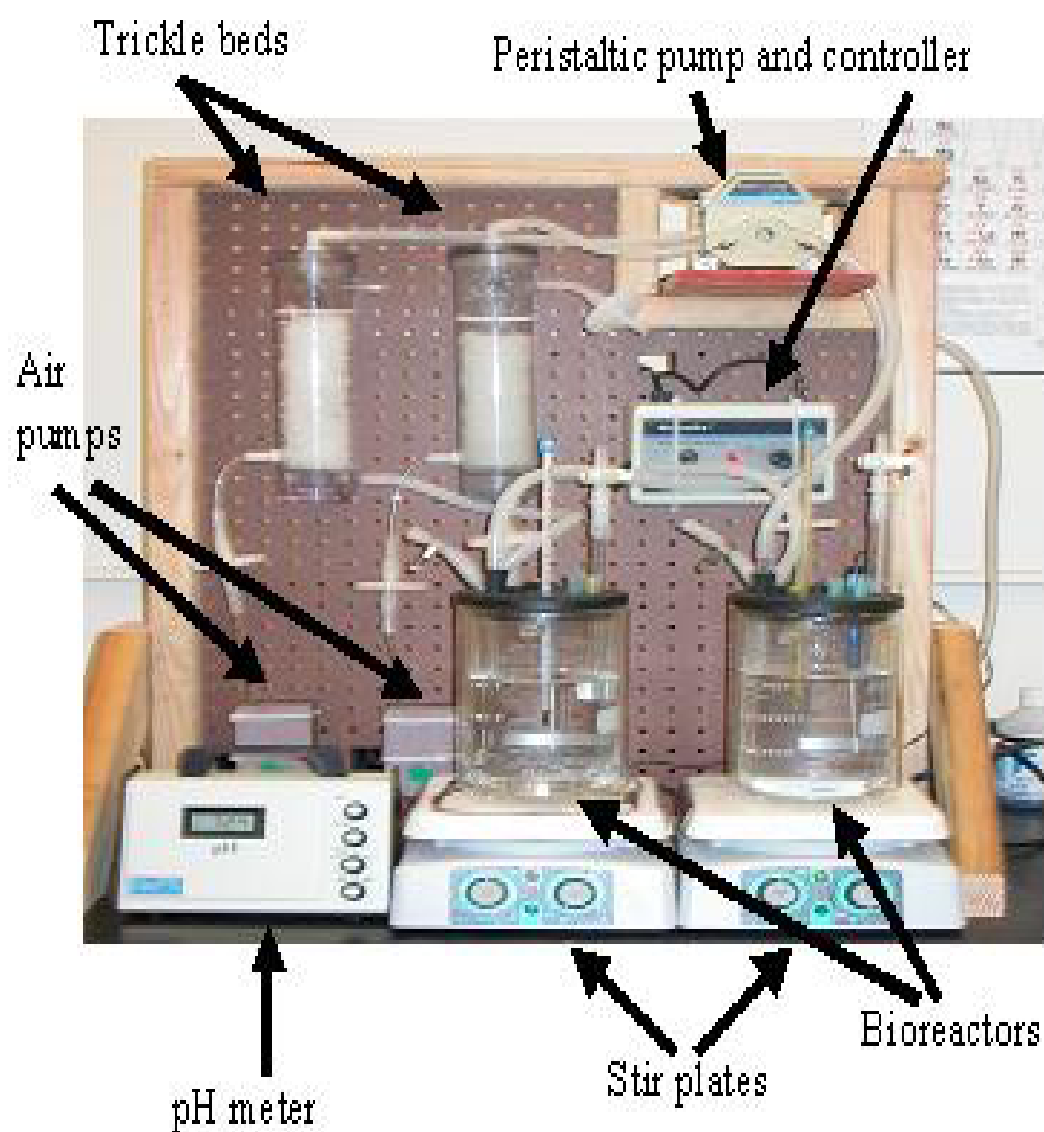
Figure 1. Biofilm reactors for water treatment testing biogenic Mn oxide produced by *L. discophora.*
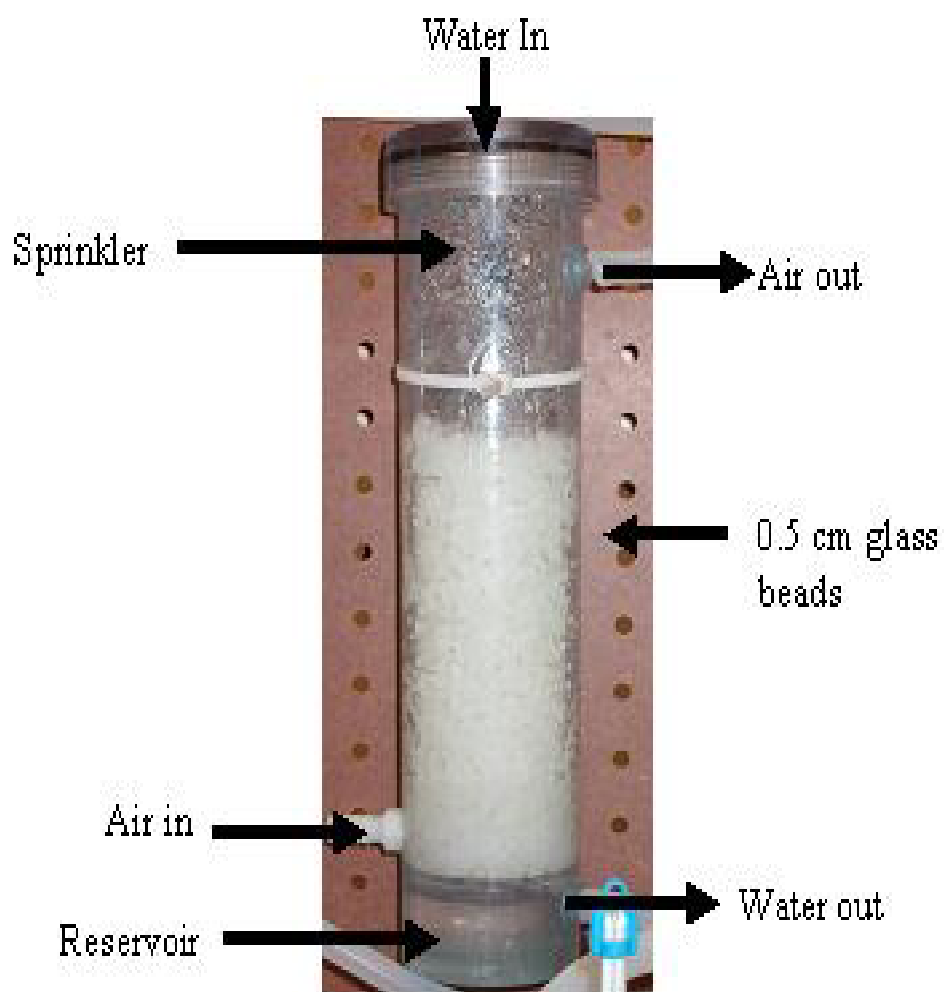
Figure 2. Close-up view of the pilot-scale biofilm reactor. This bioreactor is a continuous-flow fixed-film system with plastic beads for packing. Glass beads will be used for final design.

References:

Adams, L. F. and W. C. Ghiorse (1986). "Physiology and ultrastructure of *Leptothrix discophora* SS-1." *Arch. Microbiol.* **145**: 126-135.

Bartlett, R. J. 1988. Manganese redox reactions and organic interactions in soils. p. 59-73. *In* R. D. Graham, R. J. Hannam and N. C. Uren (ed.) Manganese in soils and plants. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Catts, J. G. and D. Langmuir (1986). "Adsorption of Cu, Pb and Zn by δMnO_2: applicability of the site binding-surface complexation model." *Appl. Geochem.* **1**: 255-264.

Nelson, Y. M., L. W. Lion, W. C. Ghiorse and M. L. Shuler (1999). "Production of biogenic Mn oxides by *Leptothrix discophora* SS-1 in a chemically defined growth medium and evaluation of their Pb adsorption characteristics." *Appl. Environ. Microbiol.* **65**(1): 175-180.

Nelson, Y.M., L. W. Lion, M.L. Shuler, and W. C. Ghiorse (2002) "Lead adsorption to mixtures of biogenic Mn oxides and Fe oxides." *Environ. Sci. Technol.:* **36**, 421-425.

Nelson, Y.M. and L. W. Lion (2003) "Formation of biogenic manganese oxides and their influence on the scavenging of toxic trace metals." Book Chapter in <u>Geochemical and Hydrological Reactivity of Heavy Metals in Soils</u>, H.M. Selim and W.L. Kingerly, Eds., CRC Press.

Nriagu, J. O. and J. M. Pacyna (1988). "Quantitative assessment of worldwide contamination of air, water and soils by trace metals." *Nature* **333**(12): 134-139.

Sunda, W. G. and D. J. Kieber (1994). "Oxidation of humic substances by manganese oxides yields low-molecular-weight organic substrates." *Nature* **367**: 62-64.

U.S.EPA (1993). Cleaning Up the Nation's Waste Sites: Market and Technology Trends, U. S. Environmental Protection Agency.

USEPA (1984). Hazardous Waste Sites: Description of Sites on Current National Priorities List, U.S. Environmental Protection Agency/ Remedial Response Program.

Zhang, J., L.W. Lion, Y.M. Nelson, M.L. Shuler, and W.C. Ghiorse (2002) "Kinetics of Mn oxidation by *Leptothrix discophora* SS-1". *Geochimica et Cosmochimica Acta*: **65**, 773-781

------------ BLANK PAGE -------------

A FAULT-TOLERANT FLIGHT CONTROLLER DESIGN

Xiao-Hua (Helen) Yu

Electrical Engineering Department
Cal Poly State University

------------- BLANK PAGE -------------

**VLSI DESIGN AND TEST WITH INTERNET-BASED
SOFTWARE AND HARDWARE TOOLS**

William Ahlgren

Electrical Engineering Department
Cal Poly State University

VLSI DESIGN AND TEST WITH INTERNET-BASED SOFTWARE AND
HARDWARE TOOLS

Report for the period August-September 2003

by

William Ahlgren
Electrical Engineering Department

Abstract
Very Large Scale Integrated Circuit (VLSI), a.k.a System on Chip (SOC), design refers
to microelectronic systems incorporating from 1 to 1000 million transistors on a single
silicon chip. The design and test of systems of such complexity is possible only with the
use of computer-aided design software tools and computer-controlled test hardware.
Access to these tools by universities and small businesses is limited, not only by the
initial investment required, but also by the recurring cost of operation and maintenance.
It is proposed to develop a plan for an Integrated Systems Design and Test Consortium
that will exploit Internet 2 technology to make the necessary tools more widely available.
The Consortium will complement the capabilities of the MOSIS organization
(www.mosis.org), which provides for fabrication of VLSI systems, but not design and
test.

Background
The EE Department is working to incorporate state-of-the-art VLSI design tools from
Cadence Design Systems into the curriculum. These tools are not easy to install and
maintain, raising questions about how they can be supported in a resource-limited
environment. A suggestion was made by Craig Schultz of ITS that the support costs be
shared by several CSU campuses, by housing the tools on one campus, and making them
available system-wide through use of the Internet. This proposal was put before Cadence
Design Systems. Although there was some interest in supporting Web-based tools,
Cadence noted that making it work is a non-trivial project, that could not be supported on
a short time-scale. They suggested defining a project that could undertake to realize the
desired capability, and perhaps more. The purpose of this document is to report progress
toward defining such a project.

Progress
The project has two parts, internet-based design, and internet-based test.

Internet-based design We began by familiarizing ourselves with prior work. In the area
of internet-based design, there is a significant literature. Much if it is directed toward
software development to enable large teams of remotely located individuals or groups to
effectively collaborate together on complex design projects. Our goal is just to enable

the use of computationally intensive CAD tools physically installed on a server at one location, to be accessed and used at a remote location, through the internet. Some literature relevant to this more restricted goal is cited as References 1 to 9. We also learned that a related topic was discussed at a panel session organized by Mark Johnson of Purdue University, at the Microelectronic Systems Education Conference on June 1, 2003. The result of that meeting, was apparently not a plan for internet-based tool sharing, but only for support on tools installed on university premises.

Our goal is fairly straightforward technically, but challenging to achieve because it depends on business decisions based on profit motive. We analyzed the obstacles to initiating a demonstration project. The vendors of the CAD tools must embrace a new business model. CAD tool vendors today prefer to license their tools for installation at the customer's site, where the customer is responsible for maintenance of the tools. As we found at Cal Poly, maintenance of these complex software packages is not the trivial task that one might think, but requires considerable attention from a skilled Unix system administrator. This is a problem for many universities, small, and even not-so-small companies. An alternative, enabled by the availability of high-bandwidth internet connections, is to physically locate the tools on hardware maintained by the tool vendor, overseen by the tool vendor's system administrators. Charging for the use of the tools might then be done on a metered basis, like taking a ride in a taxi cab; or by purchasing access for fixed periods of time, similar to traditional licensing. A significant issue is data security, as competition-sensitive data would be traveling over the internet, and customers who are direct competitors might be sharing the same server at the tool vendor. Such issues can be solved by applying data security technology already available for the government, and for financial and other industries.

Running an experiment in the academic community allows these issues to be explored in a non-critical environment. Contacts were explored at UC Berkeley, seeking a partner for an experiment. A potential collaborator unexpectedly left the university, and we have not yet identified a new collaborator there. Cadence Design Systems has suggested a collaborator at San Jose State University, and will pursue that suggestion. Permissions from Cadence, the tool vendor, are required in any case. Discussions with Cadence have yielded the likelihood of near-term support for a development project on a cost-sharing basis. Further information is provided below, under "Future Plans". We have also made contact with Mentor Graphics, another tool vendor, to explore the possibility of tool donations and internet experiment permissions from them. These discussions are ongoing. We have also contacted a local internet security firm, InfoGard, which has expressed interest in teaming with us on future developments in this area.

Internet-based test There is a substantial literature on the subject that is relevant to our project. Much of it has to do with remote experimentation for distance learning. There is an extensive literature on this subject; there is even a textbook, Ref. 10, which we cite in lieu of a more comprehensive list of individual published papers. Unny Menon has publicized the fact that the Sloan Foundation has a project to develop and popularize the distance education applications, and that Cal Poly has been invited to be a beta test site. One specific development, reported in Ref. 11, is of special relevance. A remote test

facility has already been developed in Australia, very similar to what we envision for the U.S. In fact, this facility is available to international users (http://teletest.org.au). We think that the availability of this facility does not obviate the desirability of having a similar capability in the U.S., but rather validates it. In building such a facility, we would like to have the collaboration of two leading integrated-system test equipment manufacturers, Teradyne and Agilent. We have initiated contact with Teradyne, and seeking a contact person within Agilent.

Future Plans

We need to continue this study with the collaboration of industrial and university partners. We are actively working to establish contacts in several companies. We have an existing relationship with Cadence Design Systems, which has indicated that support for an initial project might be available for 2004 (next month!), especially if there is cost sharing. Cadence indicated that funds on the order of $10K might be made available from their side. We will present more information in a follow-on proposal.

References

1. L. Geppert, "IC design on the World-Wide-Web." *IEEE Spectrum*, June 1998, pp. 45-63.
2. S. Hauk and S. Knol, "Data security for Web-based CAD." *IEEE Transactions on Design Automation of Electronic Systems*, 1999.
3. N. H. Kapadia, M. S. Lundstrom, and J. A. B. Fortes, "PUNCH: a software infrastructure for network-based CAD."
4. H.-P. Wen, C.-Y. Lin, and Y.-L. Lin, "Concurrent-simulation-based remote IP verification over the internet for System-on-a-Chip designs." *ISSS'01*, October 1-3, Montreal, pp. 233-238.
5. A. de Geus, "The Web is the catalyst for design productivity." EE Times, June 18, 1999 (http://www.eet.com/story/OEG19990618S0025).
6. R. Goering, "Genedax, Avant! cut new paths for Web-based EDA." *EE Times*, June 21, 1999 (http://www.eedesign.com/story/OEG19990621S0003).
7. Anonymous, "Barcelona offers first Web-based desing service, for op-amp optimization." Semiconductor Business News, April 10, 2000 (http://www.siliconstrategies.com/story/OEG20000410S0048)
8. J. R. Lineback, "Startup aims to accelerate analog optimization with Web-based engines." Semiconductor Business News, February 28, 2000 (http://www.siliconstrategies.com/story/OEG20000228S0001)
9. M. Santarini and R. Goering, "Web-based EDA tools find limited acceptance." *EE Times*, April 18, 2001 (http://www.eet.com/story/OEG20010418S0076).
10. T. A. Fjeldy and M. S. Shur, *Lab on the Web: Running Real Electronics Experiments via the Internet*. Wiley, 2003.
11. K. Eshragian, S. Lachowicz, and S. Eshragian, "The Australian national network for tele-test facility for integrated systems." *Proc. SPIE* 4591, 22-27 (2001).

Internet Map Server for Geographic Information: Further Development

Walter Bremer

Landscape Architecture Department
Cal Poly State University

<div align="center">

Final Report
Project: 55441

Internet Map Server for Geographic Information: Further Development
Walter Bremer, Landscape Architecture Department
Cal Poly State University

Sponsor: ONR #N00014-02-1-0754

</div>

Introduction
Geographic information systems (GIS) can benefit for Internet 2 technologies for a number of reasons. First, geographic information systems tend to move and display large data sets, a natural fit for the bandwidth of Internet 2. Secondly, the nature of GIS tends to be collaborative since the use of geographic data to solve environmental problems typically involves many disciplines. Internet 2 can facilitate both the collaboration and sharing activities.

Project Results
The purpose of this grant was to continue the development of the Internet Map Server (IMS) project started in the fall of 2002. The Landscape Architecture Department purchased the hardware to develop an Internet Map Server (IMS) to use as the development machine for the construction of map projects to be deployed on the university IMS, to have an experimental machine to design new IMS interfaces, and to explore GIS collaboration. The IMS is housed in the Geographic Information System Technology (GIST) lab in the department.

This project hired two students to assistant the work of Walter Bremer during the summer of 2003 to finish the implementation of the IMS located in the Landscape Architecture Department.

The tasks for the project are as follows:
- Document the implementation of the deployment of the IMS to date
- Refine the implementation to redirect the Tomcat processes through the Apache web server
- Develop additional map projects and deploy them on the IMS site

To date these tasks have been completed and the resulting IMS is located at:

> http://larndarch.larc.calpoly.edu/gist/steelhead
> http://landarch.larc.calpoly.edu/gist/paso_study

With the completion of this project, further expansion and use of the IMS is continuing for ongoing GIS research projects and is being used in the department to serve GIS data for applied studio projects.

<div align="center">

W. Bremer - Project 55441
Page 1

</div>

A Two-tiered Installation of ArcIMS 4.0
Cal Poly Geographic Information Systems Technology (GIST) Lab

Introduction

During the summer quarters of 2002 and 2003, work was carried out in the Geographic Information Systems Technology (GIST) lab in order to install the ArcIMS application, a web-based Geographic Information System (GIS). Our work focused on achieving a two-tiered deployment of ArcIMS (see attached diagram). In this scenario, the ArcIMS application is served by two tightly coupled hosts running complementary software. This approach, which shares the overall computing requirements of ArcIMS between two machines, provides for faster serving of maps and other content. This document details the process to successfully carry out the intended installation. Unless noted below, one should follow the typical installation procedures for ArcIMS.

Software and hardware used

In this implementation, a Macintosh OS X Server host (landarch.larc.calpoly.edu, or "landarch") provides servlet and http services, while the actual ArcIMS application runs on a Sun Solaris host (gist.calpoly.edu, or "gist"). Details of the two hosts are as follows:

landarch

Operating system: Macintosh OS X Server 10.2.6
Web server: Apache 2.0.47 web server
Servlet Engine: Tomcat 4.1 servlet engine
Java Virtual Machine: Java 1.4.1 JVM

gist

Operating System: Sun Solaris 8.0
Application: ArcIMS 4.1 Geographic Information System (GIS) web application
Java Virtual Machine: Java 1.4.1 JVM

Enabling file sharing between landarch and gist

One challenge of deploying a distributed system concerns how to enable the sharing of files between two host machines. In order for a two-tiered installation of ArcIMS to function, a framework that allows both systems to access shared files must be in place. To achieve this, we used the NFS (Network File System) protocol / application.

In this setting, the directory that holds our GIS data is located on the gist machine, in the "/usr/gist" directory. While logged into gist, this directory is made available to other

systems by using the share command. The share command exports, or makes a resource available for mounting, through a network connection. While logged into landarch, a connection to these data is created by using the mount command. The mount command is used to graft a file system onto the local file structure.

To ease the administration of the NFS connection easier, two scripts are placed in the /Library/StartupItems/MountGIST directory of landarch. These scripts ensure that each time the landarch system restarts, the file mounting is automatically carried out. The following is a listing of the first file, MountGIST (this file must have the same name as the folder it resides in).

```
#!/bin/sh

. /etc/rc.common

if [ "${MOUNT_GIST}" = "-YES-" ]; then

    ConsoleMessage "Mounting gist"
    mount 129.65.38.42:/usr/gist /usr/gist
```

The second file we included was "StartupParameters.plist":

```
if [ "${MOUNT_GIST}" = "-YES-" ]; then

    ConsoleMessage "Mounting gist"
     mount 129.65.38.42:/usr/gist /usr/gist
fi
[landarch:/Library/StartupItems/MountGIST] bmwood% ls
MountGIST              StartupParameters.plist
[landarch:/Library/StartupItems/MountGIST] bmwood% more
StartupParameters.plist
    {
     Description    = "script to mount gist via nfs";
     Provides       = ("gist_nfs");
     Requires       = ("Resolver");
     OrderPreference = "None";
     Messages =
    {
         start = "mounting gist via nfs";
         stop  = "Stopping mounting gist via nfs";
    };
    }
```

Integrating Apache and Tomcat

Another challenge faced during our work was to enable communication between the Apache web server and the Tomcat servlet engine. In this installation, both the web server and the servlet engine reside on the landarch system. The task of integrating Apache and Tomcat on a OS X Server was initially more difficult than with other variants of Unix or the Windows OS since it is a new UNIX based operating system. This task was started during the summer of 2002, and by 2003 Mac OS X Server implemented the technologies to make the task more feasible.

There are several advantages to integrating Apache and Tomcat. For instance, the URL an integrated system is of the familiar and standard type, e.g., "http://www.landarch.calpoly.edu". In contrast, accessing an ArcIMS site that runs Apache and Tomcat without integrating the two will require that the URL specify the port on which Tomcat is operating, e.g. "http://www.landarch.calpoly.edu:8080/"

In addition to providing a more typical URL, a system that integrates Apache with Tomcat will most likely see gains in overall performance of serving web content. The Apache web server has been designed to serve HTTP requests, while the Tomcat server is specifically designed for the serving of java servlets. Integrating the two allows each application to be dedicated to its intended function.

The implementation uses the "mod_jk" Apache modue to integrate Apache with Tomcat. In order to inform the Apace web server to load this module, the following line is included at the bottom of the Apache configuration file, "httpd.conf"
    Include /Library/Tomcat/conf/mod_jk.conf

This directive informs Apache to load the file "mod_jk.conf" during its startup procedure. This file handles the configuration of the mod_jk module. The following is a listing of the contents of mod_jk.conf:

```
<IfModule !mod_jk.c>
LoadModule jk_module /Library/Apache2/modules/mod_jk.so
</IfModule>

JkWorkersFile /Library/Tomcat/conf/workers.properties
JkLogFile /Library/Tomcat/logs/mod_jk.log
JkLogLevel info

<VirtualHost landarch.larc.calpoly.edu>
ServerName landarch.larc.calpoly.edu

JkMount /examples ajp13
JkMount /examples/* ajp13

JkMount /webdav ajp13
JkMount /webdav/* ajp13
```

```
JkMount /tomcat-docs ajp13
JkMount /tomcat-docs/* ajp13

JkMount /manager ajp13
JkMount /manager/* ajp13

JkMount /servlet/* ajp13

</VirtualHost>
```

Summary

Performing a two-tiered installation of ArcIMS has proven successful. Since its
deployment in September 2003, the system has been performing without technical
difficulties. This work was made possible by support supplied by California Polytechnic
State University, San Luis Obispo. This example installation will assist other GIS users
deploy similar systems.

Other administration notes:

The version of Apace that comes installed in Mac OS 10.2.6 is version 1.3. We have
upgraded the landarch web server to Apache version 2.0.47. This was done using a free
download from http://www.serverlogistics.com. This site also supplied a free package to
install Tomcat. We have also upgraded the version of Tomcat that comes installed with
Mac OS 10.2.6 Server. This creates an inconsistency in the system that should be noted.
As presently configured, the "server settings" application displays options for the
configuration of Tomcat. Please disregard these functions-- they refer to the old version
of Apache that came installed with the system. The new version of Tomcat that runs on
the system is not GUI-configurable.

Some simple procedures need to be followed whenever the landarch system is restarted in
order to maintain the operation of this server. When landarch restarts, the old version of
Apache automatically starts up. In order to launch the new version of Apache, please take
the following steps:

   1) In the Server Settings application, select Internet > "stop web service"
   2) In the System Preferences application, select "apache2"
   3) Select "start"

The Tomcat servlet engine starts automatically with system restart. This is achieved using
another startup script, very similar to the one listed above that performs NFS mounting at
system restart. The script is located in the /Library/StartupItems/Tomcat directory:

```
#!/bin/sh
```

```
##
# Start Jakarta Tomcat
# Aaron Faby <aaron@yourlink.net>
##

. /etc/rc.common

CATALINA_HOME="/Library/Tomcat"
JAVA_HOME="/Library/Java/Home"

StartService ()
{
  ConsoleMessage "Starting Jakarta Tomcat"
   $CATALINA_HOME/bin/start_tomcat.sh &
}

StopService ()
{
   ConsoleMessage "Stopping Jakarta Tomcat"
  $CATALINA_HOME/bin/stop_tomcat.sh &}

RestartService ()
{
   StopService
   StartService
}

RunService "$1"
```

The script above must be accompanied by the following
/Library/StartupItems/Tomcat/StartupParameters.plist file:

```
{
 Description    = "Jakarta Tomcat";
 Provides       = ("Tomcat");
 Requires       = ("Disks", "Resolver");
 Uses           = ("NFS", "Network Time");
 OrderPreference = "None";
 Messages =
 {
 start = "Starting Jakarta Tomcat";
 stop  = "Stopping Jakarta Tomcat";
 };
}
```

Two-tiered Installation of ArcIMS in the Geographic Information Systems Technology Lab

ArcIMS Client

Web Server

ArcIMS 4.1 Server

Internet 2

OS: Mac OS Server 10.2.6
Web server: Apache 2.1
Servlet engine: Tomcat 4.1
Servlet connector: mod_jk

Files shared using NFS

OS: Sun Solaris 8.0

GIS Data

landarch.larc.calpoly.edu

gist.calpoly.edu
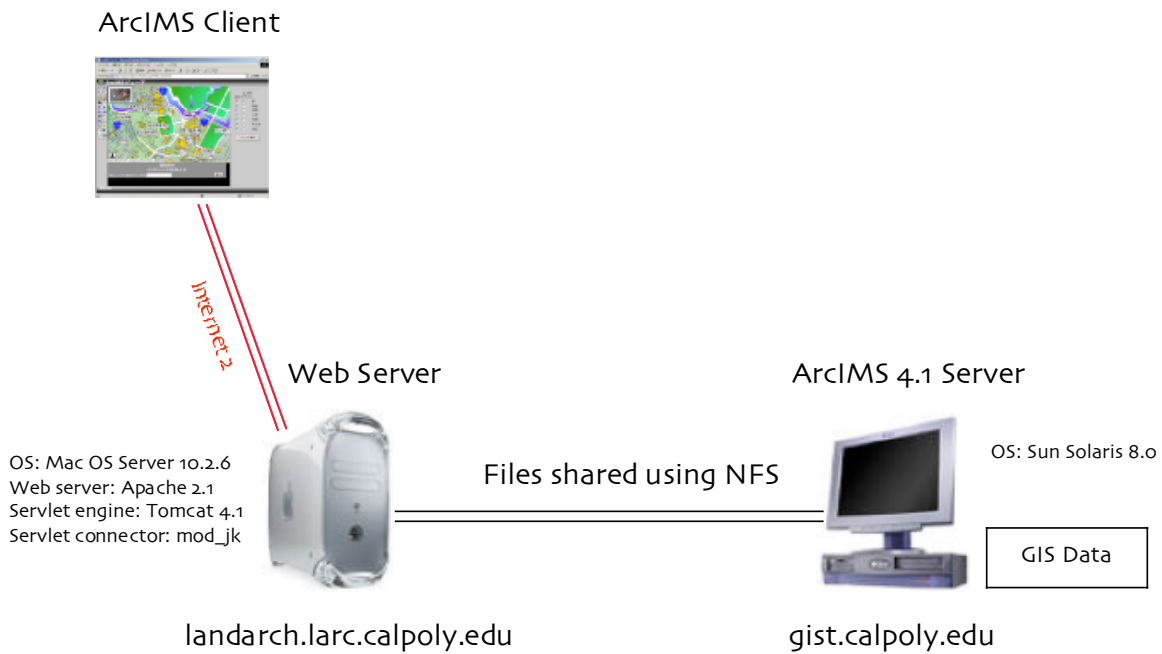
I2-Based Remote Viewing System

Fred W. DePiero

Electrical Engineering Department
Cal Poly State University

Final Report
Project: 55431
Title: <u>I-2-BASED REMOTE VIEWING SYSTEM</u>
Sponsor: Office of Naval Research (ONR)/N00014-02-1-0754

The application area for my work is an advanced type of remote viewing system. The system would provide greater flexibility than with conventional methods, such as are used in teleconferencing. Using my techniques each user could select their own viewpoint of a remote scene. However this is accomplished without installing a separate camera (and camera positioner) for each user – such an approach would not scale well for multiple users. Rather my method combines sensor data from multiple perspectives and then renders (computes) views for each user. The challenges of my work are accuracy when combining sensor data (registration) and the development of methods that require deterministic worst-case processing times. These are both important for real-time viewing. Sensor positioning and motion planning are also important to mitigate occlusion effects when rendering imagery.

This past year I improved my registration algorithms and worked on a networked implementation that includes transmission and visualization components.

Working with graduate students Kurtis Kredo and Greg Fryer, we began development of a stand-alone networked processor. The vision for this device is a low-cost ($100) general-purpose computing platform that could work together with similar units. A large number of these could be placed on a network to share the processing load.

With the inclusion of the networked transmission and visualization components as well as the algorithmic improvements, I now have a near complete, end-to-end implementation of a remote-viewing system. My setup does lack a real-time sensor ($100k) but development of the processing components is well established.

The baseline techniques that I am developing (registration) also have a number of other potential application areas, such as the automatic construction of digital elevation maps. This is of interest to potential military sponsors. Pursuing the remote viewing system seems to be a fruitful route for the present. This type of system seems to resonate well with sponsors, who seem to appreciate potential spin-off applications.

The following publications describe technical details of my work this past year. Both of these papers appeared in peer-reviewed, international conferences.

F. W. DePiero, "Deterministic Surface Registration at 10Hz Based on Landmark Graphs With Prediction," *14th British Machine Vision Conf. (BMVC2003), Norwich, UK, Sept 9-11, 2003.*

F. W. DePiero, "Surface Registration at 10Hz Based on Landmark Graphs - Benefits for a Scalable Remote Viewing System," *3rd International Conf on Visualization, Imaging, and Image Processing (VIIP2003), IASTED, Benalmadena, Spain, Sept 8-10, 2003.*

Deterministic Surface Registration at 10Hz
Based on Landmark Graphs – With Prediction

Abstract

Landmark graphs provide a means for surface registration, based on determining subgraph isomorphism to find scene-to-scene correspondences. Surface data used herein included both range and colour imagery. Images were acquired of a static scene from a moving sensor. The continuous motion allowed the sensor position to be predicted. This stabilized graph formation. Landmarks were determined using the KLT corner detector. Graph structure was established using nodes (landmarks) and edges that agreed well with predicted locations. Subgraph matching was approximated using the LeRP algorithm. A 6 DOF rigid transformation including translation and rotation was found via Horn's method. Test results on real and synthetic images indicate that a substantial speed improvement is possible, with greater determinism than ICP, while maintaining accuracy. Tests incorporated relatively large image displacements, spanning up to 30% of the sensor FOV for the image stream. Mean absolute errors of 0.6-0.9% FOV were found. Mean compute rates were ~10 Hz with standard deviation ranging 6-9%, for an image size of 200x200. Tests were run on a 900 MHz PC. 32 test trials are reported, with comparisons against a fast version of ICP.

1 Introduction

Future viewing systems could permit multiple users to explore scenes of remote locations with greater flexibility than is possible today. Rather that provide each user with their own video camera, positioner and transmission channel, the techniques described herein support a different approach. Images for each user would be rendered from a set of 3-D surface data, based on a selected viewpoint. Surface registration is the process of aligning the 3-D data into an expansive, contiguous set – beyond the field of view of a single sensor image.

Tele-presence viewing and immersion VR require 3-D sensing, registration, transmission and visualization technologies, some of which are unavailable today for real-time systems. For example a system using small 320x200 sensor images acquired at 25 Hz (motion picture rate), would need to acquire and process nearly $2 \cdot 10^6$ points/sec. For perspective, some state-of-the-art commercial 3-D sensors are within an order of magnitude of this rate – such as the Perceptron LASAR at 325k pts/sec and others [31]. Real-time registration is another key technology, in addition to sensing, that is very demanding and still requires further study. A rate of 400k pts/sec is reported herein.

Given a sensor that acquires both surface location and colour, and given a real-time registration capability, it would become possible to render live images from viewpoints that are distinct from sensor locations. This would provide a substantial degree of flexibility for remote viewing systems compared to current systems. Applications such as a 'television with a joystick' would become possible. For example in a sports broadcast, some viewers might choose to watch the hands of a golfer, others the ball, others the whole putting green. Granted more complex scenes might have a high level of occlusion, despite multiple sensors.

For tele-immersion, two such views could be computed, one for each eye. Another application is tele-medicine, where this sort of system could provide useful flexibility. For example if a field technician positioned a range & colour sensor over a patient's wound, then a remote doctor could examine the injury. Furthermore, if the doctor's viewpoint were graphically presented to the sensor technician, then the technician could anticipate the doctor's viewing needs – in terms of standoff or locations, for example. This would permit a much more efficient viewing experience for the doctor. In

another approach, the technician could possibly be replaced by a robot, which would use the doctor's viewpoint as a basis for path planning when positioning the sensor.

All of these advanced viewing systems rely on surface registration that can be accomplished in a fast and deterministic fashion. The fundamental reason that registration is required is because sensors such as laser range finders (or simple video cameras) are all line-of-sight devices. Hence either multiple sensors or multiple images (from a moving sensor) would typically be required to form a complete set of surface data across an entire scene. The focus of this effort has been on static scenes with a moving sensor.

2 Approach

The long-term goal of this research is to pursue a technique that performs view registration automatically at rates ~25 Hz, with relatively large image sizes and large sensor displacements. This performance goal targets remote viewing applications, with rapid sensor motion.

To achieve fast and deterministic processing, iterative [1] [8] [13] [17] [18], compute intensive [10], or random [6] approaches were avoided. Note that reported methods often do not separate the steps of determining corresponding points and determining the transform [8]. This limits compute speed. In the new landmark-graph approach these steps have been kept separate, and are implemented in a non-iterative fashion. This is an important distinction. Another difference is that correspondence between the data sets is determined only for select landmarks as opposed to traditional ICP, for example [1], where an entire range image is used in the cost function. This improves processing speed, but it does potentially limit accuracy because not all the scene data is used to find the transform.

The landmark-graph method does not rely on an affine motion model [25][24][14]. This permits relatively large disparities to be accommodated. For example, some of the image sequences in [14] appeared to contain a displacement of ~1/5% of the FOV between subsequent images. Tests scenes documented herein ranged 2.5% to 5%. Thus the landmark-graph approach is an alternative to a stereo-based approach [29] for the large disparities.

Stability of the landmark-based registration is achieved by maintaining consistent inter-landmark (3-D) geometry, which is verified via the LeRP [4] subgraph-matching algorithm. This is in contrast to approaches such as [14] which provide robustness based on checks of deviation in the path of each individual feature, but that do not enforce a specific geometrical structure (attributed graph) between features.

A few remaining points distinguishing the landmark-graph approach from other reported methods: this method does not require photometric normalization as with [25]. It is also not reliant on a smooth surface assumption as with [2]. No particular assumptions regarding scene content (such as planar surfaces [28]) are made – but that the KLT landmark locator [12] by employable. (KLT responds to corners or sharp prominences in the range data). And finally, the method does solve for all 6 DOF of the translation and rotation, as opposed to [24][27].

2.1 Notation and Processing Summary

The following notation is used to describe the processing and representation of a stream of surface data. The stream is composed of a sequence of range and colour images, indexed by i = 0,1,2...

   I. F_i, Sensor coordinate frame for i_{th} scene.
   II. (R_i, C_i) Range & color images acquired at location F_i.
   III. L_i, Set of landmarks found in R_i (w/rt F_i).
   IV. G_i, Graph formed from landmarks L_i.
   V. T_i, Coordinate transform relating F_i to F_0.
   VI. V_0, Graph associated with all landmarks for entire image stream.
   VII. V_i, Predicted subgraph of V_0, approximating G_i.
   VIII. V_i', and G_i' are matched subgraphs.

The following steps are used to process each (R_i, C_i) sensor input, to yield a coordinate transform T_i.

   I.   Acquire new sensor images R_i & C_i.
   II.   Predict V_i based on V_0 and on estimate of sensor motion.
   III.   Find landmarks L_i in range image R_i. Reject unstable Li.
   IV.   Form G_i using L_i, mimicking structure of Vi, (both nodes and edges).
   V.   Compute attributes for G_i, using R_i & C_i.
   VI.   Use LeRP algorithm to match G_i to V_i, the resulting subgraph mapping gives the L_i to L_0 correspondences.
   VII.   Find transform T_i via Horn's method, using the L_i to L_0 correspondences.
   VIII.   Check residual error from Horn. Remove outliers and recompute T_i.
   IX.   Repeat

Table 1: Notation and Processing Steps for Landmark Graph-Based Registration

2.2 Determining Landmark Location

Important performance goals for landmark detection and localization are: accuracy of location, speed and determinism of computations, and stability. Here, stability of detection refers to the consistent appearance of a given landmark, despite small shifts in sensor position, or despite sensor noise.

Figure 1. Example of scene landmarks found in two range images. Landmarks are determined via the KLT corner detector. These images contained a 5-degree shift.

Some reported techniques use invariant features that involve curvature classification, moments, or spherical harmonics [18]. These kinds of features rely on local smoothness for proper surface characterization. Jump discontinuities [19] violate this smoothness assumption. Local peaks may be adequate in some applications, such

as registering aerial range imagery [23]. However, this simple form of detection may not be sufficient to handle generic scenes.

Several methods for landmark detection were investigated in this effort. It was observed during experimental trials that ridge curves [20][21] appeared to be relatively stable. However, this approach was not pursued in light of challenges in parameterizing their shape. The use of neural nets, operating on a binary image of jump discontinuities [19] was also studied. However, the corner detector developed for the KLT algorithm (Kanade-Lucas-Tomasi) [12] appeared to be more tolerant to sensor noise, in terms of node stability and accuracy, compared to the neural networks.

KLT computes the eigenvalues associated with a covariance matrix of intensity gradients. The gradients are computed over a small 5x5 window. A 'corner' is associated with jump discontinuities that occur (ideally) along two perpendicular axes. Corners give rise to two large eigenvalues for the covariance matrix. An implementation from [7] was integrated into the system. Note that only the feature detector of [12] was used, not the tracking part of the KLT algorithm.

It may be appropriate to mention that care should be exercised when finding landmarks near jump discontinuities. A landmark should be selected on the nearer side of the jump. The scene location at the far side of the jump – along the sensor line-of-sight at the discontinuity – is inherently unstable. A ridge along a cylindrical surface is another example of an inherently unstable landmark. Using the KLT corner detector, and selecting the nearer surface, avoids these kinds of instabilities.

Post processing was used to improve the accuracy of landmark locations. This was necessary in part because the [7] implementation incorporates a multiresolution image pyramid, for improved speed. This does however appear to limit accuracy. The post processing also assured that the landmark was on the nearer surface of a jump discontinuity.

In the post-processing step, a small neighbourhood surrounding the original location (from KLT) was examined to find the sharpest corner. The nearby pixel with the largest fraction of distant neighbours was selected as the new landmark. The more distant neighbours were identified using a simple 'jump' threshold – in the range image. Hence, the measure of the local sharpness was simply the fraction of more distant neighbours surrounding the landmark. This attribute proved useful for colouring the landmarks (described next) and for eliminating unstable landmarks.

The KLT corner detector reacts to any variation in shape that occurs within a window of pixels. It is possible that within a given window, a foreground object might have a simple straight edge, while a background region might present additional curvature. The net result can make the KLT detector misfire. This is an undesirable result as the straight edge of the foreground object is not a corner, and hence is an unstable location. The local sharpness measure helped eliminate these cases. A threshold of 60% was applied to the count of more distant pixels. This eliminated the straight edge cases. This general sort of refinement of the results of KLT is also discussed in [14].

2.3 Determining Landmark Attributes

Node colours were represented by a 2x1 vector quantity. The components included hue from the colour image and the measure of local sharpness, described above.

Due to imperfect alignment of the sensor's range and colour images, it is desirable to provide improved robustness when combining data from these inputs. The examination of errors when using an inter-camera calibration model revealed problems near jump discontinuities (as to be expected). A simple approach of median filtering the hue values in the (mapped) neighbourhood of the landmark appeared beneficial.

2.4 Determining Graph Structure and Edge Attributes

Previously implementations of landmark graphs [23] used a Delaunay triangulation [5] [16] to determine which node locations were linked by an edge. For applications with

large standoffs, such as with aerial range imagery, the triangularization may be an acceptable approximation. However, the assumption of planarity was not appropriate for the close-range scenes used in this investigation.

As described in Table 1, a graph V_0 is maintained for all landmarks in the sensor data stream. As such, V_0 could potentially grow without bound – depending on the sensor trajectory. This is undesirable from the standpoint of subgraph matching. A limit on graph size is preferred for computational speed. To provide this size limit a subgraph Vi, of V_0, is predicted. The subgraph Vi contains nodes that are expected to be within the field of view of the sensor, given the sensor's previous position and the velocity estimate. V_i is formed using V_{i-1} and then adding or removing portions of V_0.

A graph G_i must also be generated for each new set of sensor data (Ri & Ci). The structure of G_i is established in a manner that mimics the structure of Vi as much as possible. For this, an edge in G_i is introduced between two landmarks L_i^a and L_i^b if

$$| L_i^a - L_0^x | < t \text{ and } | L_i^b - L_0^y | < t$$

Where L_0^x and L_0^y are a pair of landmarks in G_0 that are connected by an edge, and 't' is some appropriate threshold of distance.

This can lead to the introduction of 'redundant' or 'parallel' edges in G_i, compared to the structure of Vi. As such, graphs G_i tend to be supersets of ideal matching subgraphs. This is acceptable because the subgraph matching algorithm works well in the presence of this type of structural difference. See test trials described in [3].

The attribute, or colouring, for graph edges is simply the Euclidean distance between nodes. Distances were computed in world coordinates, not just a pixel distance, making the edge colouring tolerant to standoff changes. This formed an object-centered representation that could be compared without first aligning the range images.

2.5 Approximating Subgraph Isomorphism

Noisy sensor data introduces some fundamental limitations to the accuracy and stability of landmarks. This limits the similarity of scene graphs. Varying occlusion with different sensor viewpoints also limits the similarity of scene graphs. For these reasons, graphs made from real sensor data tend to be imperfect representations. Since the graphs are imperfect, an exact method of determining subgraph isomorphism has limited use – and consumes inappropriate compute time. Hence using an approximate method of graph matching is a reasonable approach.

Subgraph matching is accomplished using the LeRP Algorithm [4]. Also see Appendix 1. LeRP approximates a subgraph isomorphism via a deterministic procedure, based on the comparison of length-r paths. The LeRP algorithm yields a set of corresponding locations in the two input scene images, from which the absolute orientation may be found.

Node and edge colours are compared during the matching process. As these colours are continuous quantities, a threshold on colour differences was applied.

2.6 Determining Absolute Orientation

Horn's method [9] was used to determine absolute orientation. This reveals all 6 degree of freedom in the rigid transformation. It operated on the corresponding landmarks identified by the graph matching operation.

After an initial pass with Horn's technique, the residual error of each (3-D) landmark location is computed. The presence of outliers is checked via a simple threshold and the associated landmarks are removed and the transform recomputed. The process of finding the transform and checking consistency was repeated (fewer than) **F** times, for **F** landmarks. Hence the effort in this stage of the processing is bounded by O(**F²**). A minimum of 3 pairs of corresponding points is needed to Horn's method. Typically no

outliers were removed, but occasionally some were present and this step helped improve accuracy.

3 Summary of Computational Effort

Each of the processing steps requires worst-case effort that has a polynomial bound. The effort described in Table 2 assumes: **NxM** range images and **F** landmarks. **D** is the mean degree of landmark graphs and **Q** is the total number of edges.

Step	Sub-Step	Effort
Find Graph Vi	Predict Sensor Motion	O(Constant)
	Find Vi from Vi-1, and V0	O(F)
Find Landmarks Li	Find KLT Corner Features	O(N M)
	Refine Landmark Locations	O(F)
	Reject Unstable Landmarks	O(F)
Find Graph Gi	Find Structure Mimicking Vi	$O(QF^2)$
	Colour Nodes	O(F)
	Colour Edges	$O(F^2)$
Match Graphs Vi and Gi	LeRP algorithm	$O(F^3D^2)$
Find Coordinate Transform Ti	Use Horn's Technique	O(F)
	Remove Outliers & Recompute	$O(F^2)$
Update Graph V0	Refine Landmark Locations	O(F)

Table 2: Computational Effort for Landmark Graph-Based Registration

Processing effort for the graph matching is on the order of $O(F^3D^2R)$, where **F** is the number of nodes, **D** is the average degree. The parameter **R** is actually a weak function of **F** (see [4]) but was set to a constant in all tests reported herein.

4 Testing and Results

As shown in Table 3, two types of scene data were used. The 'Sensor' data was acquired using a structured light sensor, similar to [3], that was built in-house. The device includes a mechanical positioner, permitting it to collect the 3-D data set. The sensor also includes a colour camera [30].

The in-house sensor is relatively slow, relative to commercial 3-D range cameras. Furthermore, it is these state-of-the-art range cameras that are driving the goals for the new registration technique. Hence range images from the sensor were acquired and then stored for use in testing. Point clouds generated by the sensor were resampled to form scene imagery, as shown in Figure 4.

The 'Synthetic' test scenes were generated using models of blocks, with ray tracing calculations to determine each pixel of the scene. Gaussian noise was added to the range and hue images during testing, spanning 2% of the intensity range.

Tests with synthetic scenes benchmarked accuracy for both sensor translation and rotation. Translational steps were 0.1 inches – 3% of the 3.6" sensor FOV. The total shift over an image sequence was 22% of the FOV. Rotational steps were of 1 degree spanned a total range of 5 degrees. A total of 23 unique synthetic scenes were used in testing. Synthetic scenes were reused used on multiple test trials, by adding random noise. Scenes scanned with the in-house sensor had translational steps of 2.5% (of an 8" FOV) with an overall span of 30% FOV. Objects scanned with the real sensor had a

fairly consistent standoff. Hence the synthetic images appeard somewhat more cluttered.

4.1 Benchmarks Against a Fast-ICP Algorithm

An ICP algorithm was used for comparison purposes [1][18]. ICP was implemented with a simplex optimisation routine [15]. To reduce the effort of evaluating the cost function, a simple image difference was used. In each evaluation the point cloud was transformed and a range image was formed. The mean absolute difference in pixels was then computed. This avoids the step of having to search for the closest point in the point cloud. It does introduce error, however, because the closest 3-D point might actually be found via a lateral shift to an adjacent pixel. To improve compute speed only the grey level difference with the current pixel was considered. This type of resampling is also described in [24].

Another variation on ICP used herein maintained a fixed number of iterations. This permitted more direct comparisons between ICP and the landmark-graph method, as both algorithms were deterministic. Exactly 200 iterations of ICP were run in each trial.

Another speed optimisation step for ICP involved limiting the region of interest used in the cost function comparisons to areas near landmarks. This somewhat intertwined the implementation of the two techniques. This approach is being investigated as a potential means to create a hybrid technique that combines ICP with the landmark-graph method.

4.2 Test Results

Table 3 summarizes a total of 32 trials. Each 'trial' consisted of one registration operation on a given pair of images. Percent error was computed via the absolute mean displacement error, and then expressed relative to the sensor field of view. Rotational errors are also given as absolute means. All means were computed over the entire image sequence, including initial transient.

-	-	#	Length	Trans. Error		Rotation. Error		Rate: Mean +/- Std. Dev. (Hz)	
Scene	Type	Trials	Stream	Graph	ICP	Graph	ICP	Graph	ICP
Synthetic	Rotation	5	6	0.6%	0.6%	1.7^0	1.0^0	10 +/- 9%	0.13 +/- 20%
Synthetic	Trans.	16	9	0.1%	0.2%	1.7^0	0.7^0	10 +/- 6%	0.14 +/- 21%
Sensor	Trans.	11	12	0.6%	0.6%	1.1^0	0.3^0	10 +/- 6%	0.14 +/- 6%

Table 3. Results of test trials indicate that the accuracy of the landmark graph approach is comparable to Fast-ICP, while providing a substantial speed improvement. ICP was run with 200 (a constant number of) iterations. Mean absolute error values are reported. All images were 200x200; tests run on a 900MHz PC.

As seen in Table 3, the landmark graph approach rivals the accuracy of Fast-ICP, while executing much faster, at ~10Hz. The landmark-based approach also provides a greater degree of determinism than ICP as indicated by the standard deviation of the processing time. Fast-ICP was run with 200 iterations and the images for all test trials were 200x200. Tests were run on a 900MHz PC.

A reason ICP suffers variations in processing rate despite the use of a fixed number of iterations is due to the simplex-based optimisation [15]. The simplex method will use a varying number of evaluations of the cost function, depending on the type of simplex movement ('contraction' versus 'flip'). Note that other optimisation techniques also possess this type of computational variation per iteration, such as Hooke-Jeeves.

The landmark-graph approach also suffers some variations in processing rates, as shown. This is due to varying numbers of landmarks (F in Table 2). The degree of this variation can be mitigated somewhat, for example if the system processes an image stream with substantial changes to scene content, by adjusting detection threshold of KLT. Dynamic means of threshold adjustment are of interest in future studies. Limits on **F** are also possible when computing Graph Vi. See Table 2.

Figure 4. Landmarks and scene graph for range data acquired with a structured light sensor.

5 Conclusions and Future Studies

Reported results are encouraging. The landmark-based approach is able to achieve rates ~10Hz for 200x200 images on a single PC. 1% accuracy appears comparable to ICP and to other reported techniques. Yet landmark-graphs provide higher speed and greater determinism than ICP. The method also yields 6 DOF alignment parameters and can process image streams with relatively large jumps in sensor position (5%). The ability to handle large changes in sensor position not only allows for fast sensor motion, but also indicates that precise predictions of sensor motion are not critical for accurate registration.

The landmark-graph approach does have its share of challenges. Most significant is the computation of landmarks via a method that is both fast and stable. The KLT feature detector appears to work well. It would be beneficial to improve the rotational invariance. A circular neighbourhood might help in this regard. This is an on-going area of investigation.

Stable placement of graph edges was a challenge in a previous implementation [23], but the predictive approach that mimics graph structure of V0 seems much better.

Future extensions could also include marrying the new technique with ICP in a post-processing step. This would permit more scene data to enter into the final transform calculations than just the landmarks. ICP could begin iterations using the alignment parameters derived via landmark graphs. A fixed number of iterations could also be used for determinism. A goal here would be to target improved rotational accuracy and it might be best to just search over those 3 DOF. As with the implementation here, the region of interest used in cost function evaluations for Fast-ICP can be restricted using the image regions near corresponding landmarks.

A general improvement to the landmark-graph approach is planned in the near future. This involves a post-processing step, where the graph V0 is grown in size after each new image is aligned. This is an O(F) operation, see Table 2. Growing V0 in an on-line fashion would permit extended regions of surface data to be incorporated into a single contiguous data set. The implementation reported here is more suited for repeated sensor scans over a small area, from differing views.

Future work is also planned for an integrated visualization subsystem to provide a complete real-time remote viewing system.

6 Acknowledgements

References

[1] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, IEEE Trans. Pattern Analysis and Machine Intelligence, 14 (2) (1992) 239-256.
[2] B.K.P. Horn and J.G. Harris, Rigid Body Motion from Range Image Sequences, CVGIP: Image Understanding, 53 (1), pp.1-13, 1991.
[3] F. W. DePiero and M. M. Trivedi, "3-D Computer Vision Using Structured Light: Design, Calibration and Implementation Issues," AIC, vol.43 (1996).
[4] F. W. DePiero and D.W. Krout, LeRP: An algorithm using length-r paths to approximate subgraph isomorphism, Pattern Recognition Journal, 24 (2003) 33-46.
[5] O. Faugeras, 3-D Comp. Vision, A Geometric Viewpoint, MIT Press, Cambridge, MA, 1993.
[6] M. Fischler, R. Bolles, Random Consensus: a paradigm for model fitting with applications in image analysis and automated cartography, Communications of the ACM, 24 (1981) 381-395.
[7] Birch, Implementation of the Kanade-Lucas-Tomasi Tracker
http://vision.stanford.edu/~birch/klt/.
[8] B. Luo and E.R. Hancock, Structural graph matching using the EM algorithm and singular value decomposition, IEEE Trans. PAMI, 23 (10) (2001) 1106-1119.
[9] B.K.P. Horn, Closed-form solution of absolute orientation using unit quaternions, J. Optical Society of America A, 4 (4) (1987) 629-642.
[10] A. E. Johnson and S. B. Kang. "Registration and integration of textured 3-D data." Image and Vision Comput., 17, 135-147, '99.
[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.
[13] C. Kapoutsis, C. Vavoulidis, and I. Pitas, Morphological iterative closest point algorithm, IEEE Trans. On Image Processing, v. 8, no. 11, pp 1644-1646, Nov. (99).
[14] T. Tommasini, A. Fusiello, E. Trucco and V. Roberto. Making good features track better. in Proceedings IEEE Computer Society Conference on Computer Vision Pattern Recognition, 1998, pp. 145-149.
[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, NY, 1988.
[16] J.R. Shewchuk, Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator, Proc. First Workshop on Applied Computational Geometry, Philadelphia, Pennsylvania, pages 124-133, ACM (May 1996).
[17] S. Hsu, Multiple-view constrained video registration and its applications, Workshop on Video Registration, IEEE Computer Society, Vancouver Canada, (July, 2001).

[18] G. C. Sharp, S. W. Lee, D. K. Wehe, ICP registration using invariant features, IEEE PAMI, 24 (1) (2002) 90-102.

[19] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice-Hall, NJ, 2001.

[20] J. T. Kent, K. V. Mardia, J. M. West, Ridge curves and shape analysis, British Machine Vision Conference, University of Edinburgh (1996).

[21] X. Pennec, N. Ayache, J-P. Thirion, Landmark-based registration using features identified through differential geometry, Handbook of Medical Imaging, I. N. Bankman ed., 499-513, Academic Press (Sept 2000).

[22] A. E. Johnson, Surface landmark selection and matching in natural terrain, IEEE Computer Vision and Pattern Recognition, v. 2, 413-420 (2000).

[23] F. W. DePiero, "Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results," Proc. 1st Intl. Symposium on 3-D Data Processing, Visualization and Transmission (3DPVT), (Padova, Italy), June, 2002.

[24] A. E. Johnson and A. Miguel San Martin, Motion Estimation from Laser Ranging for Autonomous Comet Landing, IEEE Computer Vision and Pattern Recognition, v. 2, 413-420 (2000).

[25] J. Shi and C. Tomasi. "Good Features To Track." *Proc. Computer Vision and Pattern Recognition* (CVPR'94), pp. 593-600, 1994.

[27] E. Noirfalise, J.T. Lapresté, F. Jurie and M Dhome, Real-time Registration for Image Mosaicing, Electronic Proc. of The 13th BMVC, University of Cardiff, 2-5 September 2002.

[28] Gille Simon and Marie-Odile Berger, Real time registration of known or recovered multi-planar structures: application to AR, Electronic Proc. of The 13th BMVC, University of Cardiff, 2-5 September 2002.

[29] A. Fusiello, V. Roberto, and E. Trucco. Efficient stereo with multiple windowing. In *CVPR*, pages 858–863, 1997.

[30] F. DePiero, Structured Light Range Sensor, http://www.ee.calpoly.edu/~fdepiero.

[31] J. Hancock, et. al., Active Laser Radar for High Performance Measurements, Proc 1998 IEEE Intl. Conf on Robotics & Automation, Leuven, Belgium, May 1998.

Appendix – LeRP Algorithm for Approximating Subgraph Isomorphism [4]

 Main Routine
Input: Graph **G** with nodes **g$_i$, 0<=i<N$_G$** and Graph **H** with nodes **h$_k$, 0<=k<N$_H$**
Output: Mapping **m()**, that gives **h$_k$ = m(g$_i$)**.
Steps:
I.   Compute powers of adjacency matrices **A^R** and **B^R** for graphs **G** and **H**
II.   **beta$_{peak}$[][] = *find_best_beta*(G,H,A^r,B^r)**
III.   Clear node-to-node mappings
IV.   For each **L, 0<=L<minimum(N$_G$,N$_H$)**
    A.   Let **peak = 0**
    B.   For each unmapped node **g$_i$**
    C.   For each unmapped node **h$_k$**
        1.   Verify consistency of mapping **g$_i$** to **h$_k$** given current **m()**
        2.   **rho = 0**
        3.   For each mapped edge **e$_{ij}$**
            a)   lookup associated edge **e$_{kl}$** where **l=m(j)**
            b)   **beta = *compare*(i,j,k,l)**
            c)   **gamma = *compare*(j,j,l,l)**
            d)   **rho = 1 – (1-rho)(1-beta)(1-gamma)**
        4.   Next **j**
        5.   **alpha = *compare*(i,i,k,k)**
        6.   **rho = 1 – (1-rho)(1- alpha)(1- beta$_{peak}$[i][k])**
        7.   If **rho>peak** Then
            a)   **g$_{peak}$=i**
            b)   **h$_{peak}$ =k**
            c)   **peak=rho**
        8.   End If
    D.   Next **k**
    E.   Next **i**
    F.   If **peak=0** Then GoTo **END**
    G.   Let **m(g$_{peak}$)=h$_{peak}$**
V.   Next **L**
VI.   If **(L=N$_G$)** and **(L=N$_H$)** Then **G** *is ISOMORPHIC to* **H**, refer to mapping **m()**.
VII.   Else a subgraph isomorphism exists between **G** and **H**, refer to mapping **m()**.
VIII. END

Function: find_best_beta(G,H,A^r,B^r)
    A.   For each node **g$_i$**
    B.   For each node **h$_k$**
        1.   For each edge **e$_{ij}$**
        2.   For each edge **e$_{kl}$**
           a)   **beta = *compare* (i,j,k,l)**
           b)   Save **beta$_{peak}$[i][k]=beta** if maximal for nodes i,k
        3.   Next **l**
        4.   Next **j**
    C.   Next **k**
    D.   Next **i**
    E.   Return **beta$_{peak}$[][]**

 Function: compare(i,j,k,l)
I.   For **1<=r<=R**
    A.   If **a$_{ij}$$^{(r)}$!= b$_{kl}$$^{(r)}$** Then Break
II.   Next **r**
III.   Return **(r/N)^2**

Surface Registration at 10Hz Based on Landmark-Graphs: Benefits for a Scalable Remote Viewing System

Abstract – *Real-time surface registration is a key technology for the development of future remote viewing systems. An architecture for a video distribution system supporting multiple users, with individual viewpoint selection, is suggested. The approach would provide a transmission bandwidth independent of the number of users, for scalability. The proposed architecture uses a method of surface registration based on landmark-graphs. Results from 141 test trials on synthetic scenes indicate that a mean absolute positioning accuracy under 1% of the sensor field of view is possible. The mean rate for registration was 10Hz, with a standard deviation under 10%. Tests were benchmarked on a 900MHz PC. The sensor images were 200x200 pixels and contained both range and color imagery.*

1. Flexible Remote Viewing Systems

The goal of this research is to further methods of surface registration, for the enhancement of remote viewing systems. Current viewing capabilities such as TV or teleconferencing are quite limited by restrictions in viewpoint as each user is fed the same view. Furthermore, the selection is restricted to discrete camera signals.
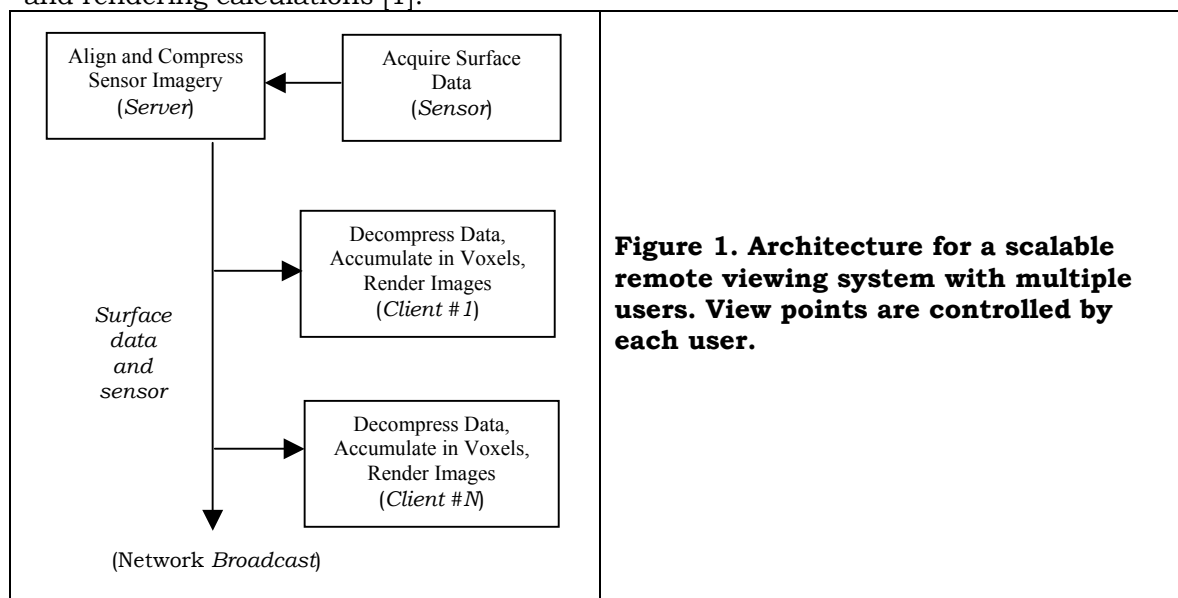
Improved remote viewing systems should provide viewpoint flexibility for multiple users. Preferably, this should be done without simply introducing a camera with pan & tilt for each user and without increasing the transmission bandwidth in proportion to the number of users.

Such improvements may be possible, given an ability to do real-time surface registration. This refers to 'stitching together' sections of a scene that have been acquired by sensor(s) from different vantage points. This permits a large contiguous set of surface data to be constructed, as a basis for rendering remote views.

Accomplishing registration in real-time means that the alignment calculations must be completed at the rate of sensor acquisition, thus permitting immediate use of the sensor data for remote viewing. Voxel-based rendering could then provide imagery with an arbitrary viewpoint.

Given the real-time registration capability new approaches to video distribution become possible. See Figure 1. The server acquires new sensor images, and then computes an alignment relative to previous inputs. By transmitting the new sensor data to clients along with alignment transformations, the rendering operations may then be off-loaded to client machines.

This approach permits each client to have an independent viewpoint. It also means that the bandwidth of the transmission is determined by the sensor(s) only, not by the number of users. The method also offloads considerable effort, by not centralizing all the processing and rendering calculations [1].

Figure 1. Architecture for a scalable remote viewing system with multiple users. View points are controlled by each user.

A visualization system should provide rapid response to user requests for new viewpoints. The proposed architecture is well optimized in this regard, as the viewpoint request and subsequent rendering are all local to the client machine. This makes the rendering frame rate and response to pan and tilt view changes all independent – and not limited by - the sensor data rate or the transmission rate.

2. Potential Applications
.

Applications such as a 'television with a joystick' would become possible, given the ability to perform real-time registration. This would support a broadcast transmission to many users, each with an independent viewpoint. For example with a sports broadcast, some viewers might choose to watch the hands of a golfer, others the ball, others the whole putting green. Scenes with an individual golfer would be amenable to this sort of remote viewing system. More complex scenes (such as a crowded street) could have a prohibitive level of occlusion, despite multiple sensors. For applications with tele-immersion, two such views could be computed, one for each eye.

Another application area is tele-medicine. A scenario is proposed here that is more flexible than just the transmission of individual medical scans. Rather, more interactive modes of observation are envisioned. For example if a field technician positioned a sensor over a patient's wound, then a remote doctor could examine the injury. Furthermore, if the doctor's viewpoint could be graphically presented to the sensor technician, then the doctor's viewing needs could be better anticipated.

In another remote-viewing scenario a robot could use the doctor's viewpoint as a basis for path planning and sensor positioning.

Awareness of another person's viewpoint is pre-attentive knowledge, when interacting directly. However, this knowledge can be lost in a remote-viewing scenario. Means to graphically present a remote user's viewpoint may be a useful feature for advanced systems.

3. Areas of Investigation

All of these advanced viewing scenarios rely on surface registration. The fundamental reason that registration is required is because sensors such as laser range finders (and even simple video cameras) are line-of-sight devices. Hence either multiple sensors or multiple images (from a moving sensor) would typically be required to form a complete set of surface data across an entire scene. Figure 2 illustrates the line-of-sight nature of a range sensor. The 2nd image has been rendered from a viewpoint that was offset from the sensor, revealing missing surface data.

Approaches for registration and visualization need to be deterministic and computationally tractable for real-time implementation. Methodologies in these areas are the focus herein. Also, this study is restricted to cases with static scenes that are scanned by a moving sensor.
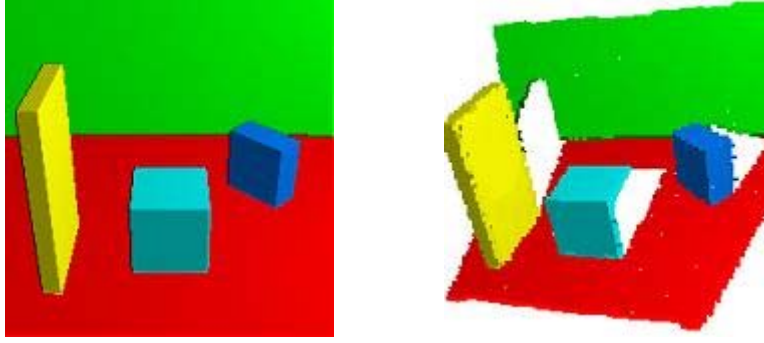
Figure 2. Sensor image (above, simulated) and a scene rendered from another distinct viewpoint. This illustrates the line-of-sight nature of range sensors. A low-resolution voxel array was used to store surface data.

4. The Challenge of Surface Registration

Surface registration is the process of determining the six DOF that describe changes in sensor location between a pair of input images. The goal here is to track changes in sensor location as the device is moved continuously across some arbitrary scene. The landmark-graph approach reveals sensor motion based entirely on an analysis of scene content – using no auxiliary sensors or alignment targets.

Work in registering range data dates back to random approaches such as RANSAC [2] and iterative methods have been widely studied [3]. However, non-deterministic methods such as these are not preferred for real-time implementations. Robust methods that are computationally intensive have also been reported [4] but may not be able to achieve high frame rates. Other methods that track features [5] assume small image displacements and then use an affine motion model to describe local scene changes. The assumption of small displacements limits sensor velocity.

Some real-time methods have also been recently proposed [6][7]. However a direct comparison to these works cannot be made, as these rely on either a fixed camera position (rotation only) [6] or on an assumption of a particular type of scene content [7].

Reported methods typically do not separate the steps of determining corresponding points and determining the transform [8][3]. These steps are kept separate for landmark-graphs thanks to the LeRP algorithm for approximating subgraph isomorphism [9]. This is an important distinction with respect to computational efficiency.

5. Surface Registration and Remote Viewing With Landmark Graphs

Stability for the landmark-graph is provided via the similarity of inter-landmark geometry, which is verified via a subgraph-matching algorithm. This is in contrast to approaches such as [5] which provide robustness based on checks of deviation in the path of each individual feature, but that do not enforce a specific geometrical structure (attributed graph) between features. See Figure 3.

The result of the graph matching processing step is a pair of subgraphs with identical structure (in terms of nodes and edges). The pair of subgraphs also has attributes that match to within specified tolerances. As such, a rigid transformation may be computed between the landmark correspondences given by the matching subgraphs.

The following notation is used, to describe the processing and representation of an image stream. The stream is composed of a sequence of sensor images, indexed by i = 0, 1, 2...

- F_i, Sensor coordinate frame for i_{th} scene.
- (R_i, C_i) Range & color images acquired at F_i.
- L_i, Set of landmarks found in R_i (w/rt F_i).
- G_i, Graph formed from landmarks L_i.
- T_i, Coordinate transform relating F_i to F_0.
- V_0, Graph associated with all landmarks for entire image stream.
- V_i, Predicted subgraph of V_0, approximating G_i.

The world coordinate frame for the voxel array is aligned with F_0. Registration calculations are based on comparisons between the i_{th} sensor location, F_i, and the initial location, F_0.

In a remote viewing system based on landmark-graph registration, the server could execute the following steps:

1) Acquire new sensor image.
2) Predict V_i based on V_0 and motion estimate.
3) Find landmarks L_i in range image R_i.
4) Form G_i using L_i, mimicking structure of Vi.
5) Compute attributes for G_i, using R_i & C_i.
6) Use LeRP algorithm to match G_i to V_i, the resulting subgraph mapping gives the L_i to L_0 correspondences.
7) Find transform T_i via Horn's method, using the L_i to L_0 correspondences.
8) Compress R_i & C_i and broadcast to clients, along with T_i.
9) Update landmark positions L_0 and attributes stored in V_0. Grow V_0 using any new territory exposed in G_i.
10) Repeat

The client could execute these steps:

1) Receive R_i & C_i along with T_i. Decompress sensor imagery.
2) Accumulate R_i & C_i into voxel array using the T_i transform.
3) Repeat.

The client would also continuously render scene images, based on the current voxel array content. This could be done asynchronously; at whatever rate the client platform can manage.

Previous work by this author with landmark-graphs restricted analyses to individual pairs of sensor images, not to image streams [10]. Stream processing is more appropriate for the continuous sensor movement. With an image stream, prediction may be used, as in [5]. Results of the landmark-graph approach, including prediction, are superior to those

previously reported by this author [10]. More information on LeRP, the graph matching technique is available in [9]. As LeRP is a relatively new algorithm, pseudo-code is included in the appendix.

6. Transmission Subsystem

Given that range data is available in addition to standard intensity images, and given the alignment data, there are new opportunities for image compression for the transmitted sensor data. Because sensor data is in the form of images, some simple variation on standard image compression techniques may be useful for the remote viewing system.

For example, the coordinate transforms T_i and T_{i-1} could be used to warp the images C_{i-1} & R_{i-1} to approximate the current images C_i & R_i. An image difference operation could then provide better compression over a method such as MPEG thanks to the warping operation that would make subsequent images more similar. Note the receiver would of course have to perform an un-warping operation.
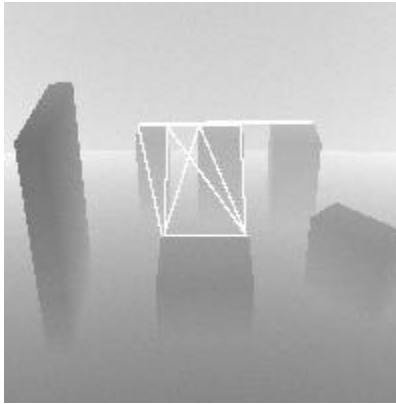

Figure 3. Landmark graph

7. Visualization Subsystem

In the proposed architecture, the client machine is tasked with accumulating range data and rendering user images. This offloads computations from the server side, making for a more balanced load. This also facilitates each user having their own viewpoint.

The method of shear-warp ray casting [11] is proposed for rendering. This method introduces a shear offset between adjacent layers effectively giving the voxel array a parallelogram shape. A projection of voxels then occurs along rows and columns of the array. This sort of projection is much more efficient than ray-tracing, for example. Projections are performed back-to-front, relative to the user viewpoint. The warp operation restores proper image aspect ratio.

The compute performance demands on client processors in this architecture may be somewhat beyond the capability of today – depending on sensor data rates and image size. However, less expensive memory, faster general-purpose processors, and voxel visualization boards [12] may all contribute to meeting these increased demands, soon.

Choices of using a voxel array, and shear-warp, were driven by the use of 3-D point clouds of sensor data and the need for real-time processing. The voxel array is well matched to the storage needs of the 3-D data points. Shear-warp then provides efficient means for rendering.

Despite the simplicity of a point cloud approach, it may have some advantages over methods that use a polygonal surface representation [13]. Consider a situation with the sensor being swept back and forth over a static scene. As new 3-D points are acquired and

aligned, a simple algorithm may be used to accumulate the data into the array – for example, just replacing the old points with new ones. Alternatively some type of averaging color values (hue) could be used when accumulating data. In contrast to this, consider a polygon-based approach. The polygons output from a sensor subsystem would have to be continuously merged to avoid unbounded growth of the surface description [13]. Such recombination and merging could be challenging in real-time. The voxel-based approach avoids this sort of problem.

8. Testing and Results

This is an on-going effort and the results of the registration with prediction are currently the main focus of investigation. Additional results documenting the effect of compression are under study.

The test suite included cases with both rotational movement and translation. Both real and synthetic sensor data has been included. Zero mean Gaussian noise was added to the synthetic sensor images.

The mean absolute position error is given as a percentage of the sensor field of view. The number of pixels across the sensor and voxel array was the same in these tests. Hence the percent error in position indicates the amount of misregistration expected in the voxel array. See Table 1.

Reports of accuracy and computational rate are given in Table 1, for both the landmark-graph approach and for a 'fast-ICP' method [10]. The fast-ICP method used a simple image difference, rather than point-by-point search for correspondence. It also ran with a fixed number of iterations (200) to yield a deterministic algorithm that is more directly comparable to the landmark-graph approach.

	Translation Synthetic Scenes	Rotation Synthetic Scenes	Translation Real Scenes
Mean Absolute Error For LG	0.1%	0.7^0	0.6%
Mean Absolute Error For ICP	0.2%	1.1^0	0.6%
LG Rate Mean +/- Std. Dev.	10 Hz +/- 9%	10 Hz +/- 7%	10 Hz +/- 6%
ICP Rate Mean +/- Std. Dev.	0.13 Hz +/- 22%	0.14 Hz +/- 19%	0.14 Hz +/- 6%

Table 1. Test results for surface registration demonstrate a faster rate and greater determinism for landmark-graphs, compared to fast-ICP.

Test results in Table 1 show relatively low errors under 1% of the sensor field of view. These mis-registration errors result in a blurring of the surface data accumulated in the voxel array. Hence these error rates of are considered good. Figure 4 shows a relatively crisp image, after the accumulation of 10 registered sensor images.

The landmark-graph method was benchmarked to be ~70x faster than ICP. Landmark-graphs also provide better determinism, see standard deviations on processing rates. These factors make the landmark-graph approach superior for a real-time system.

The processing rates are given for a 900MHz PC. Although the rates are considered good relative to other reported methods, these would still need to increase for a broadcast system. Also note that the sensor image size was only 200x200 pixels. The new method does seem promising, nonetheless, given the modest compute platform.

AVI-format video clips are available for download [14]. The clips contain images rendered during the testing discussed below. The still image in Figure 4 is from one of these sequences.
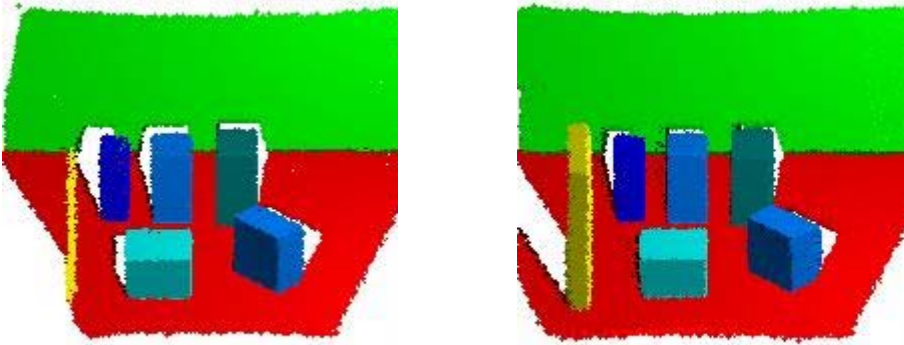

Figure 4. Rendered image from voxel array after initial sensor image, and after the accumulation of 10 images (2nd). Note the new portions of the scene encountered after all 10 images are accumulated. Also note the reduction in the missing data (white areas).

9. Conclusions and Future Studies

Test results for the landmark-graph method of surface registration appear to yield relatively crisp imagery, with registration errors under 1%. The technique could provide the basis for a new means for distribution of surface data in a remote viewing system. Such a system could support multiple users and would be a scalable architecture. Opportunities for sensor image compression are superior to standard image streams because of the registration data, which could be used to align sensor images prior to compression.

Lossy compression methods will degrade the voxel data and the final user images, as will sensor noise and registration errors. To help mitigate some of the degradation a median operation could be performed on the voxel array. This step would retain the three most recent contributions to a voxel, and use the median of the three for rendering purposes. This and other possible post processing steps are underway.

An outstanding issue in the design of the proposed architecture has to do with the introduction of new users. If surface transmissions are underway when a new client accesses the broadcasts, then the new client's voxel array will not match the state of other clients, nor of the server. Hence some means of voxel refresh would likely be required. One possibility is to provide a secondary, non-real-time transmission from the server to the clients for this purpose. The secondary transmission might consist of only filled voxels (to reduce data rates).

10. Acknowledgements

11. References

[1] VisServer by SGI, http://www.sgi.com.

[2] M. Fischler, R. Bolles, Random Consensus: a paradigm for model fitting with applications in image analysis and automated cartography, *Communications of the ACM*, 24, pp. 381-395, 1981.

[3] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, *IEEE Trans. PAMI*, 14 (2) 239-256, 1992.

[4] A. E. Johnson and S. B. Kang. "Registration and integration of textured 3-D data." *Image and Vision Comp.* 17, 135-147, 1999.

[5] T. Tommasini, A. Fusiello, E. Trucco and V. Roberto. Making good features track better. In *Proc IEEE Computer Society Conference on Computer Vision Pattern Recognition*, pp. 145-149, 1998.

[6] E. Noirfalise, J.T. Lapresté, F. Jurie and M Dhome, Real-time Registration for Image Mosaicing, *Electronic Proc. of The 13th BMVC*, University of Cardiff, September 2002.

[7] G. Simon and M.-O. Berger, Real time registration of known or recovered multi-planar structures: application to AR, *Electronic Proc. of The 13th BMVC*, University of Cardiff, September 2002.

[8] B. Luo and E.R. Hancock, Structural graph matching using the EM algorithm and singular value decomposition, *IEEE Trans. PAMI*, 23 pp. 1106-1119, 2001.

[9] F. W. DePiero and D.W. Krout, LeRP: An algorithm using length-r paths to approximate subgraph isomorphism, *Pattern Recognition Journal*, 24, 33-46, 2003.

[10] F. W. DePiero, "Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results," *Proc. 1st Intl. Symposium on 3-D Data Processing, Visualization and Transmission* (3DPVT), (Padova, Italy), June, 2002.

[11] P. Lacroute, and M. Levoy, Fast Volume Rendering Using a Shear-Warp Factorization of the Viewing Transformation, *Proc. SIGGRAPH*. ACM, Orlando, FL, pp. 451-458. 1994.

[12] Voxel visualization board, by TeraRecon, http://www.rtviz.com.

[13] Heckbert, P., and Garland, M. Survey of polygonal surface simplification algorithms. Tech. Rep. CMU-CS-95-194, Carnegie Mellon University, 1995.

[14] DePiero, AVI Clip showing results of L-G Registration, http://www.ee.calpoly.edu/~fdepiero.

Collaborative Research in Next-generation Embedded Microprocessor Designs

Diana Keen

Computer Science Department
Cal Poly State University

Internet 2 Final Report
Project Title: Collaborative research in next-generation embedded microprocessor designs
Project Director: Diana Keen
College and Department: Computer Science, College of Engineering

Abstract

Wireless communication protocols and mobile devices are on the verge of ushering in a new era in information technology. As communication protocols evolve, however, their computational needs exceed both the performance and energy budgets of conventional embedded systems. We propose to address this problem with a multi-institution, multi-disciplinary team with strengths in microprocessor design, digital signal processing, communication protocols, coding theory, circuit design, and education.

The focus of our work will be to design embedded microprocessors that approach the efficiency of special-purpose circuits used for signal processing. The purpose of the Internet2 grant was to enhance the collaborative efforts of UC Davis and Cal Poly through the use of enhanced video conferencing capabilities.   During this effort, we purchased and tested videoconferencing equipment, applied and received NSF funding, produced a workshop publication, and submitted a conference publication that is current being reviewed.

Equipment Review

In the quest for an ideal remote collaboration environment, we tested several pieces of videoconferencing equipment. I will enumerate the pieces of equipment and describe the advantages and disadvantages that we found.

1. **PolyCom ViaVideo camera for the desktop** –
Advantage: This gave an instant enhancement to the picture. The resolution was much better than the cheap camera previously used, allowing the group members to see my facial expressions as I listened and participated.

Disadvantage: For some unknown reason, it virtually stops transmitting after being connected for 20 minutes. Even with excellent help from ITS support, we were unable to figure out why. It has been tracked to the camera software, however, since the network is perfectly fine (the image I receive from Davis has no such problems). In addition, all I need to do is turn off the camera and back on and reconnect, and the picture is fine again. I tried out the new version of the camera, and the same problem occurred. PolyCom claims it is network problems, but I remain unconvinced.

2. **PolyCom robot camera** –
Advantage: The camera is amazing. The difference in the meeting was instant. The first day I had it, I did not have a true speaker phone yet, but I felt like I was so much in the meeting that I kept trying to talk to the screen. The undergraduate that was with me in the meeting was very amused that I did this not once but three times. This just shows how the full screen view with fairly good resolution was successful at making me feel like I was there.

I can control the camera from my side and view the portion of the group I want to. Alternatively, we can have the camera track the voices and automatically switch to the person

(people) talking. If two people close to each other are talking, it will even zoom out slightly so that both are in the picture rather than constantly going from one to another. The algorithm is quite good.

Disadvantage: I have found no keyboard interface to controlling the camera. This means that if I put the image on full screen, I can not control the camera, because the controls disappear. The alternative is using audio tracking. The meeting is held in a small portion of a lab at UC Davis, and the camera has problems with people sitting against walls. The sound bounces off of the wall, and the camera then points to the spot on the wall rather than the person. This is not so much a fault of the camera as the environment, and controlling the machine myself would be the solution. All they need is a keyboard interface, and it would be excellent.

3. **School-provided speaker phone**

Advantage: I can participate more because I do not need to use the phone every time I want to say something. The delay in picking up the phone made me less likely to participate fully in the discussions.

Disadvantage: The speaker phone can only send or receive, not both at the same time. This works in a very organized meeting, but when I am a single person on one side with a group on the other, I often try to say something and it does not go through at all because there is noise on the other side. This makes me a much more passive participant in the meetings. When I really need to, I grab the phone and say something. I could use the microphone in NetMeeting, but the delay is much longer than traditional phone lines, so we use traditional phone lines for our meetings.

4. **Tablet PC**

Advantage: I could draw on the tablet rather than drawing on the screen, giving the potential for increased participation in such events.

Disadvantage: It took a long time over the summer for the tabletPC to work with everything else. In addition, the robot camera has such good resolution that I could finally view the whiteboard easily. I have not yet been able to fully test out the TabletPC and will be continuing to do so in the next few months.

Research publication / funding

The project has been successful at its goals of funding and publication. We received an NSF ITR grant, on which I am a subcontractor. I will be applying for REU funding so that I can pay undergraduates to work on the project. In addition, we published one workshop paper and have submitted a conference paper that is currently being reviewed.

Conclusion

This grant has enabled much better interaction during group meetings on this project, allowing me to participate more fully. It has also given me a better idea of what is important in the meetings in order to target equipment purchases better in the future. A speaker phone that can send and receive at the same time is much more important than I would have imagined. Thank you for this opportunity to enhance collaboration with UC Davis.

Video-capture, Editing and Production on I2

Franz Kurfess
Aaron Peckham
Jordan Small

Computer Science Department
Cal Poly State University

Final Report

VIDEO-CAPTURE, EDITING AND PRODUCTION ON I2

Project Director: Franz Kurfess
Computer Science Department, Cal Poly
fkurfess@calpoly.edu
Investigators: Aaron Peckham and Jordan Small
Computer Science Department, Cal Poly
apeckham@calpoly.edu, jsmall@calpoly.edu
September 2003

Abstract
Internet2@Cal Poly is a documentary created during spring and summer of 2003 that covers the projects and applications of Internet2 at Cal Poly. The documentary aims to inspire and motivate Cal Poly professors and students to pursue projects involving Internet2. It is also intended to invite students and professors from outside Cal Poly to collaborate with Cal Poly on Internet2-related projects.

Structure of Internet2@Cal Poly
The documentary is broken down into the following sections:
Section I – Introduction
A brief segment introduces Cal Poly and the surrounding city. The montage includes views of campus and Thursday Night Market in downtown San Luis Obispo.
Section II – What Is Internet2?
Dr. Franz Kurfess and Dr. Hugh Smith introduce themselves and offer non-technical definitions of Internet2. Dr. Smith appears via videoconference. The segment shows Cal Poly's data center, where high capacity lines connect Cal Poly's network to the world.
Section III – Advantages of Internet2
This section describes Internet2 advantages by highlighting example uses: Internet2 allows engineers to experiment without affecting the stability of the commercial Internet; Internet2's high capacity enables high quality videoconferencing; and over Internet2, universities can access data-intensive remote resources such as electron microscopes.
Dr. Smith describes Cal Poly's fortunate technological and geographical position. Because of its position on California's central coast, and its strong reputation in undergraduate engineering work, Cal Poly has the opportunity to work with many other institutions over very high-speed networks.
Section IV – Internet2 Champions
In this section, Dr. Smith describes the reasoning behind the creation of the Internet2 champions, and outlines the three major goals they set out to achieve. Dr. Kurfess also talks on his personal experiences as an Internet2 champion.
Section V – Cal Poly Educational Approach
This section first highlights Cal Poly's distinct approach to education: following Cal Poly's "learn by doing" motto, many classes focus on hands-on learning. As Dr. Smith explains, this focus gives Cal Poly the opportunity to build on the accomplishments of others and concentrate on real-world applications of Internet2.

Because of Cal Poly's focus on teaching, the Internet2 Champions organized small grants, known as mini-grants, to encourage Internet2-related projects. In this section, Dr. Kurfess and Dr. Smith describe the motivation for the grants.

Section VI – Internet2 Projects at Cal Poly

This segment focuses on two Internet2 projects at Cal Poly. Dr. David Gillette from the English Department and Dr. Enrica Lovaglio from the Art and Design Project are working on the Lumiere Ghosting Project. The project's goal is to create a space that participants can enter and interact with other people and images over Internet2. Because of the large amount of data exchanged between users, the project makes use of Internet2. Dr. Hugh Smith is working on a mini-grant-funded Internet2 project that studies end-to-end performance on high capacity networks. He is working in collaboration with Ixia, a manufacturer of networking equipment.

Section VII – Internet2 Course

In this segment, Dr. Kurfess describes the Internet2 course taught in the spring of 2003 and shows segments of guest lectures from the course. These talks include lectures by Ann Doyle, Rose Bowker and Johanna Madjedi.

Section VIII – Cal Poly Goes To CENIC

This segment includes clips from the CENIC conference attended by both Jerry Hanley and Dr. Franz Kurfess' Internet2 class. Dr. Thomas West, CENIC President, appears in this segment to describe CENIC's mission.

Section IX – Future of Cal Poly and Internet2

CIO Jerry Hanley provides an overview of Cal Poly's networking strategy in this segment. Mr. Hanley encourages Cal Poly teachers and students to become more involved with Internet2, and promises that he will continue to deliver the necessary bandwidth for future projects.

Section X – Conclusion

The final segment of the documentary includes inspirational audio clips from earlier interviews, accompanied by images of campus.

Video Production

Footage included in this documentary was collected during spring and summer 2003. Video footage was taken using a Canon ZR-30 from Media Distribution Services, and still photos were taken with Canon digital cameras. Unedited footage totaled over 11 hours; the edited product is approximately 23 minutes. In total, project files used more than 60GB in storage. A short experimental video teaser was assembled during spring quarter using iMovie. That video was presented to the Internet2 class during the final week of the quarter. Based on that production the investigators decided that a more advanced editing tool was necessary. The documentary was captured and edited using Final Cut Pro 3 on an Apple Macintosh dual processor 1GHz G4 computer.

Conclusion

The documentary is a good summary of Internet2 projects at Cal Poly, and serves as an excellent tool to introduce students and teachers to the topic. Internet2@Cal Poly fits well with the ongoing efforts of the Internet2 Champions to promote use of the technology and further working relationships with other Internet2 organizations.

End-to-End Performance Testing

Hugh Smith
Computer Science Department

Cal Poly State University

Internet 2 Mini-Grant /C3RP Close out Report
Project Title: End-to-End Performance Testing
Project Director: Hugh Smith

Project Number: 55411

This is the final project report for my Internet 2 Mini-grant. This project was focused on computer network performance testing. There were three parts to this project. First, was to continue to develop our relationship with Ixia Communications in the area of network performance testing. Second, was to look at the work being done in the Internet 2 end-to-end performance testing. Third, was to look at latency and throughput on campus.

The following paragraphs detail the work done on each of these subprojects.

1) Performance testing work: We configured a VPN firewall setup in the lab. This configuration includes a Cisco PIX and a Radius server. Ixia, the performance testing company, gave us a long-term loan of their newest testbox and their high-end-testing card. This card and the new testbox will allow us to test VPN configurations and much more. After configuring the new equipment we run performance tests on our VPN configuration and provided feedback on the new card to Ixia. The results of these tests were documented in Newman Chan's senior project report. This work has contributed to us developing a relationship with Network World magazine. At the time of writing this report, we have published our first product review in Network Computer Magazine and are working on a full product comparison to be published in March.

2) End-to-end performance testing: We have been working the Colorado State University in Denver on end-to-end performance testing. This has included utilizing our Ixia testbox to study latency issues between Cal Poly and CSU Denver. A publication on this work should be submitted sometime in 1Q04. In addition, we have written a report that documents the work being done by the Internet 2 End-to-End performance group.

3) CalPoly QoS: We developed a prototype Network Weather Map for Calpoly by modifying the work done by Indiana University and CENIC. This Weather Map gives real time information on the status of backbone networking devices. Our prototype is currently running in the Cisco Advanced Networking Lab. ITS has expressed interest in this project.